Undergraduate Result Prediction Using Machine Learning Algorithms

[1]Odeniyi, L. A., [1]Ibitoye, A. O., [1]Akanni, J. O., [2]Adegoke, A, [1]Oladimeji, O. O., [1]Lawal, B., [1]Adeboje, O., and [1]Adenusi, D.

[1]*Department of Mathematical and Computing Sciences, KolaDaisi University, Ibadan, Nigeria*

[2]*Department of Computer Sciences, Chrisland University, Abeokuta, Nigeria*

***Corresponding author** - Odeniyi Latifat Adeola, *latifat.odeniyi@koladaisiuniversity.edu.ng*

Abstract

Poor academic performance has discouraged and made so many students drop out of university due to National University Commission (NUC) regulation of 1.5 probational level which consequently affects list of enrolment in higher institutions yearly. Numerous researchers have worked on this over the years in order to proffer solution to this but suffer a limitation of fewer accuracy level. Hence, this research was developed to proffer solution to the limitation by developing a result prediction system for undergraduates using data mining algorithms. In this research work, 448 instances of datasets were collected which consist of students Secondary School Certificate Examination (SSCE) results, student secondary school result in five major subjects, Joint Admission Matriculation Board (JAMB) score, and Cumulative Grade Point Average (CGPA) at their first year in the university. The machine learning algorithms employed in the research were K-Nearest Neighbor, Random Forest, J-48, and Naïve Bayes from WEKA (Waikato Environment for Knowledge Analysis). From experimental results obtained, the most valid data that affects the CGPA of an undergraduate student are the combination of JAMB result with other features such as Age and other SSCE result. The accuracy result of the developed system, using the selected algorithms are as follows: K-Nearest Neighbour (81.40%), Random Forest (85.26%), J-48 (79.65%), Naïve Bayes (76.14%). Therefore, based on the result of analysis, it was discovered that Random Forest outperformed all other machine learning algorithms. The results from this research can be employed by career advisors and Educationalists to advice students as appropriate. This research has been able to use real-world data to evaluate the success of students in tertiary institutions with the use of machine learning approach.

Keywords: Undergraduate Result Prediction system, Machine Learning Algorithms, Academic performance, Data Mining.

1. Introduction

Education is one of the most important aspects of human resource development. The students' performance plays an important role in producing best quality graduates who will become great leaders and manpower for the country thus responsible for the country's economic and social development. Academic achievement is one of the major factors considered by employers in hiring workers especially for the fresh graduates Olufemi et al [1]. Therefore, performance of students in universities should be a great concern to Government and parents and not only to the administrators, academicians and educators. Thus, students have to put the greatest effort in their study to obtain good grades and to prepare themselves for future opportunities in their career at the same time to fulfil the employer's demand Saheed et al [2]. However, poor academic performance of students is one of the major problems associated with the university environment which usually led to withdrawal of student from the programme of study right from the end of first year in the University. Government efforts to improve academic performance have led to series of interventions like scholarship, training of the lecturers, providing of infrastructure and so on. Yet, the academic performance in the Nigerian universities is declining every day. Student performance prediction system has been a promising tool to solve the problem of poor academic performance among students in the universities and to deliver high quality education. Data Mining (DM) is an approach to understand and discover hidden knowledge and pattern that can help in decision making from pool of dataset. DM techniques can be used in various areas including education, health sector, banking, security, marketing and so on to enhance maximum result and efficiency Oladimeji et al [3]. Educational Data Mining is an important area of study in data mining for exploring educational data

for various academic decisions. It can be used for prediction of students' academic performance prediction and decision about students' academic progress. Recently, Machine Learning (ML) has emerged and gain popularity as a very powerful technique from computational and statistical methods used for predictions and data pattern inferences. ML can be used to solve many problems in marketing, banking, medical, industry, agriculture, and power generation, Odeniyi and Tunwase [4]. Furthermore, ML can be used in education domain especially in higher education institutions to predict undergraduate result in order to avoid loss of students. Researchers in education mostly apply ML in two tasks. The first task is a prediction task for student performance, and the second task is a classification task which classifies students into groups based on the students' profiles. In prediction task, several ML techniques were used in research. In classification techniques evaluation, many ML techniques were applied, Yacoob et al [5].

2. Review of Related Literatures

Due to high accuracy and prediction, quality data mining technique is widely used in different areas. Education sector is also enriched with the help of this technique. A number of journal and literature are available containing educational data mining. Few of them are listed below for reference.

Bhardwaj and Pal [6] conducted study on the student performance by selecting 300 students from 5 different degree colleges in India. In their study, it was found that students' grade in senior secondary exam, living location, medium of teaching, mother's qualification, family annual income, and student's family status were highly correlated with the student academic performance. Bhardwaj and Pal [7] in their research work used students' previous semester marks, class test grade, seminar performance, assignment performance, general proficiency, attendance in class and lab work to predict students' mark in their end semester. Kostopoulos et al. [8] used enrollment data to predict successful and unsuccessful students in New Zealand and he found 59.4% and 60.5% of classification accuracy while using decision tree algorithms CHAID and CART respectively. Yadav et al. [9] conducted study on the student retention based by selecting 398 students from MCA course of VBS Purvanchal University, Jaunpur, India. By means of classification they show that student's graduation stream and grade in graduation play important role in retention.

Pal [10] conducted study on the student dropout rate by selecting 1650 students from different branches of engineering college. In their study, it was found that student's dropout rate in engineering exam, high school grade; senior secondary exam grade, family annual income and mother's occupation were highly correlated with the student academic performance.

Yadav [11] conducted a study using classification tree to predict student academic performance using students' 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.

In another study, Yadav [12] used students' attendance, class test grade, seminar and assignment marks, lab works to predict students' performance at the end of the semester with the help of three decision tree algorithms ID3, CART and C4.5. In their study they achieved 52.08%, 56.25% and 45.83% classification accuracy respectively. Merceron and Yacef [13] concluded that association technique requires not only that adequate threshold be chosen for the two standard parameters of support and confidence, but also that appropriate measures of interestingness be considered to retain meaning rules that filter uninterestingness ones out.

In Navamani and Kannammal [14], evaluation was analyzed on various techniques, including Decision Tree, Neural Network, and Naïve Bayes. Mishra and Tyagi [15] compared between several techniques, including Decision Tree, Neural Network, Naïve Bayes, Support Vector Machine, K-Nearest Neighbour, and Linear Regression to find the best ML technique among them. According to reviewed paper, the most used data set for prediction are CGPA and internal assessment score. While the Neural Network and Decision Tree are the two methods that are mostly used by the researchers for predicting student's performance. Although results from these research

studies are useful for self-study planning or teacher guidance to improve student's study achievement, the particular students have to seek for help from consultants personally. Since consultants are limited in number, students found it very hard to get help to solve their personal study problems.

Baashar et al. [16] extended their research in Electrical Engineering student performance to compare prediction efficiency between Neural Network and Linear Regression techniques. The independent value in the research also used student core subject performance in semester one and three, not only semester one as in their previous study. Model performance was measured by coefficient of Correlation R and Mean Square Error (MSE). From experiments, Neural Network yields higher predicting accuracy than Linear Regression in co-efficiency of correlation, and outcomes still indicate students' core subject performance, which strongly influences other subjects' performance.

Adekitan and Noma-Osaghae [17] in Design model for predicting the student's performance based on their grade level used Decision Tree, Naïve Bayes and Neural Network for their classification and discovered that Decision Tree produces the highest accuracy rate of 77.778%, followed by the Neural Network with accuracy rate of 70.886% and the Naïve Bayes produces the lowest at accuracy rate 66.865%.

Chen et al. [18] used diverse machine learning classifiers to predict the enrollment and dropout status in the postgraduate level, they found that the locally weighted learning outstrips the other classifiers. Sweeney et al. [19] carried out a study using Data mining approach to predicting the performance of first year student in a university using the admission requirements. Hu et al. [20] proposed a system for predicting the grades of the students for the next enrollment term They applied two classes of methods: Simple Baselines and Matrix Factorization (MF) based methods. The lowest prediction error was achieved by the Factorization Machine (FM) Model of Matrix Factorization based methods.

3. Methodology

Data were collected through Google forms sent to students. The collected data attributes were: Sex, Age, five core subject result in SSCE in either WAEC or NECO, SS2 results for the five core subjects, JAMB score and CGPA at the end of first year. The flowchart, sequence diagram and Use case diagram for the research are as shown in Figures 1, 2 and 3, respectively. The collected dataset was preprocessed in WEKA (version 3.9.5) using Machine learning Algorithms.

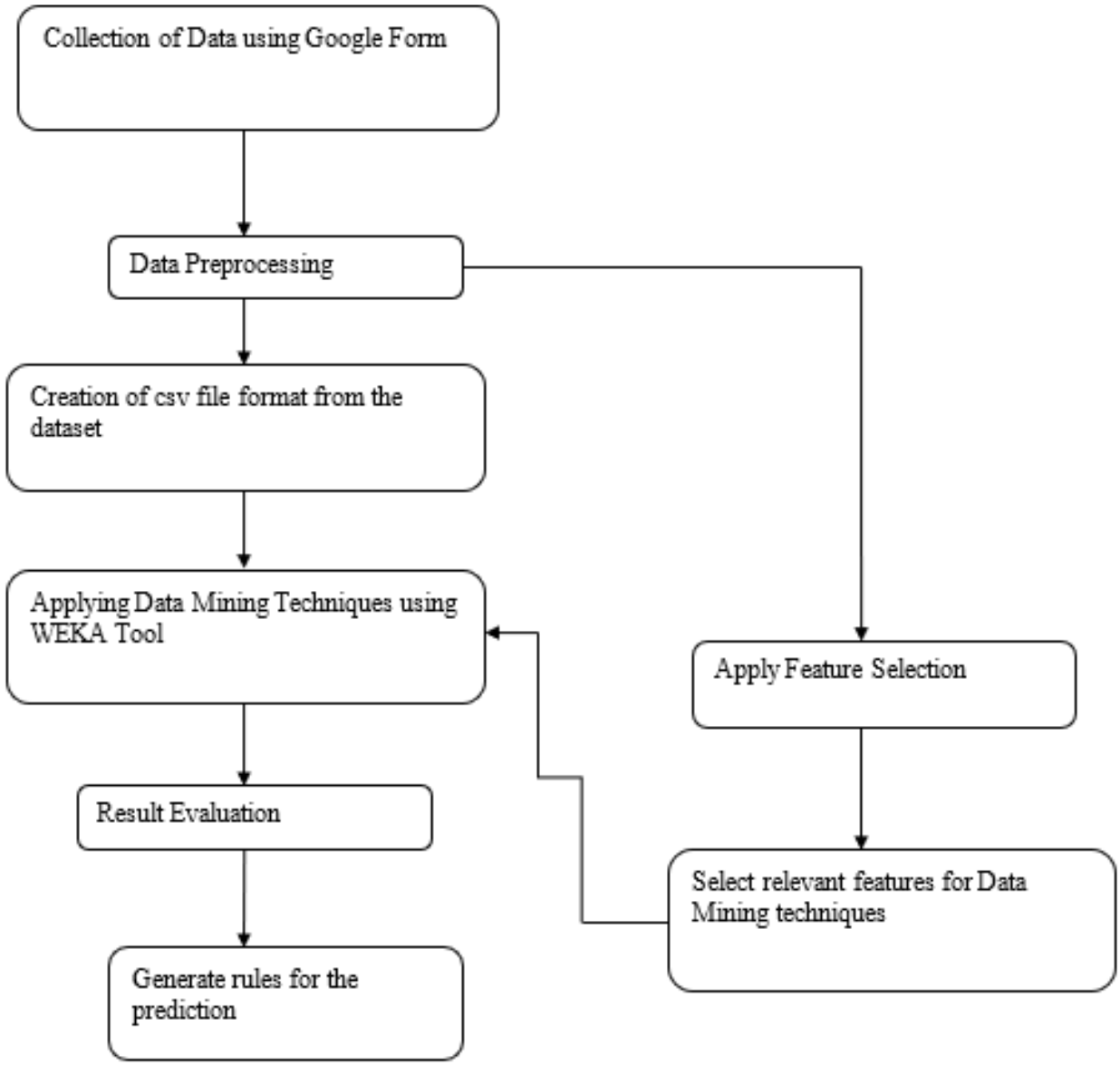

Figure 1: The flowchart showing the flow of event in the system.

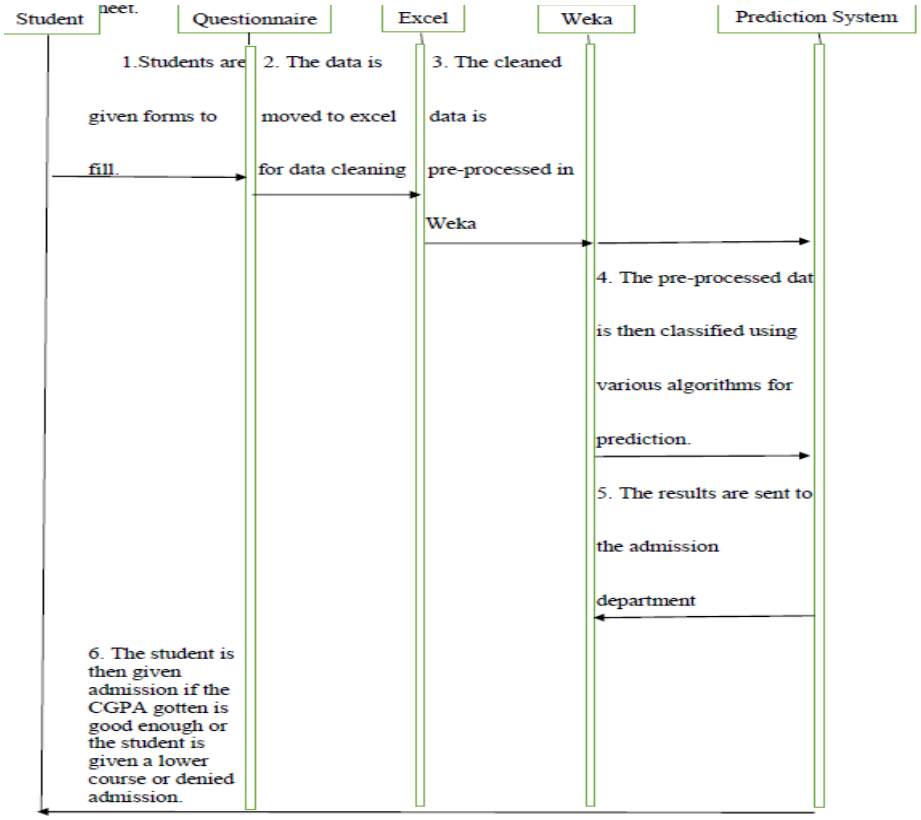

Figure 2: Sequence diagram showing the flow of operations.

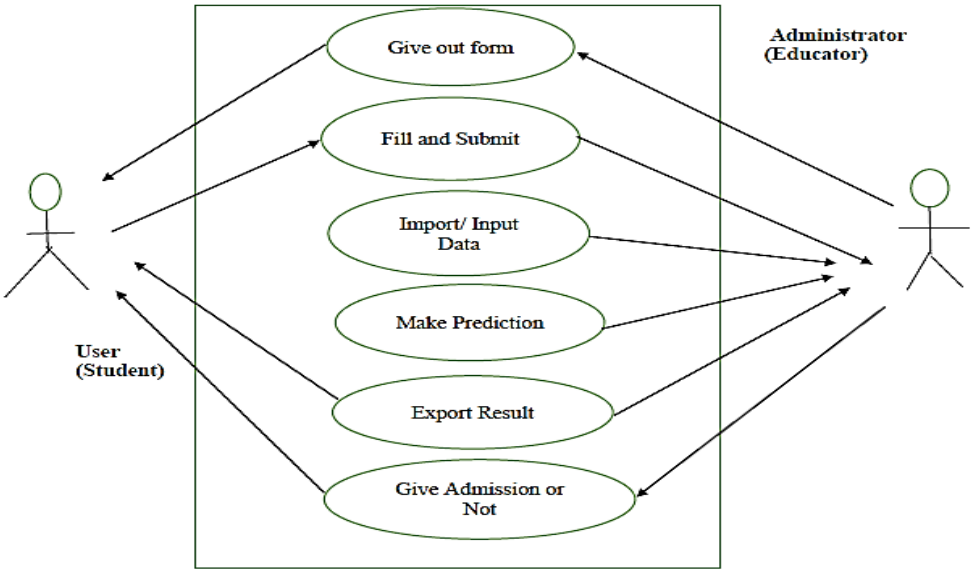

Figure 3: Use case diagram of the system.

3.1 Data Description and Preprocessing

The dataset used to pinpoint this research was gotten from undergraduate student of both private and public secondary schools of Ogun State in Nigeria. The dataset entails 448 instances in which the features include the gender, some SSCE subject results, Jamb score and SS2 results. During this stage it was discovered that the dataset was highly skewed. Hence, there is a need to alleviate the data imbalance problem. Therefore, Synthetic minority oversampling technique (SMOTE) [21] which is a technique that create synthetic data from minority samples was used.

3.2 Feature Selection

The feature selection phase entails selecting the features that are necessary for the prediction task and neglecting the irrelevant features. In this research, Information Gain and Gain Ratio algorithms coupled with ranker search method were used to rank the features of the dataset. The result of the ranking is described in Table 1.

Table 1: Summary of the Feature Ranking

Feature	Information Gain	Gain Ratio
Jamb Score	1.0302	0.2877
Age	0.735	0.2492
SSS3 Result Economics	0.6785	0.2369
SSS3 Result Biology	0.5835	0.2241
SSS3 Result Maths	0.5787	0.2167
Biology	0.5294	0.2138
SSS3 Result Civic Education	0.5025	0.2007
SSS2 Result English	0.4742	0.1965
SSS2 Result Biology	0.447	0.1929
SSS2 Result Civic Education	0.4169	0.1904
English Language	0.4083	0.1867
Mathematics	0.403	0.1857
SSS3 Result English	0.3986	0.1736
SSS2 Maths	0.3693	0.1704
Economics	0.3667	0.1577
SSS2 Result Economics	0.3077	0.1466
Civic Education	0.2655	0.1181
Sex	0.0923	0.0955

Based on the result of the feature ranking, Sex, Civic Education, SS2 Result Economics were removed from the dataset due to their feature rank as they have little or nothing to infer for this classification task.

3.3 Classification

After the data preprocessing stage, classification algorithms including Random Forest, K-Nearest Neighbour, J48 and Naïve Bayes were implemented in Waikato Environment for Knowledge Analysis (WEKA) platform. For evaluation purpose, the dataset was split into two whereby 70% of the dataset was used for the training and the remaining 30% was used for testing.

4. Results

This section showcases the result obtained before and after the application of SMOTE. The metrics used for evaluating the effectiveness of the algorithms include accuracy, recall, precision, F-measure and Mathew's correlation coefficient (MCC) [21] which are standard evaluation metrics for machine learning. Performance of the Algorithms used before and after feature selection was displayed in Tables 2 and 3 while the corresponding histograms was shown in Figures 4 and 5 below.

Table 2: Performance of the Algorithms before Feature Selection

	Accuracy	Precision	Recall	F-measure	MCC
Random Forest	83.16%	0.839	0.832	0.830	0.791
KNN	80.70%	0.796	0.807	0.797	0.751
NB	77.54%	0.784	0.775	0.776	0.720
J48	78.25%	0.781	0.782	0.775	0.726

Table 3: Performance of the Algorithms after Feature Selection

	Accuracy	Precision	Recall	F-measure	MCC
Random Forest	85.26%	0.865	0.853	0.850	0.819
KNN	81.40%	0.803	0.814	0.806	0.766
NB	76.14%	0.759	0.761	0.757	0.697
J48	79.65%	0.793	0.796	0.793	0.741

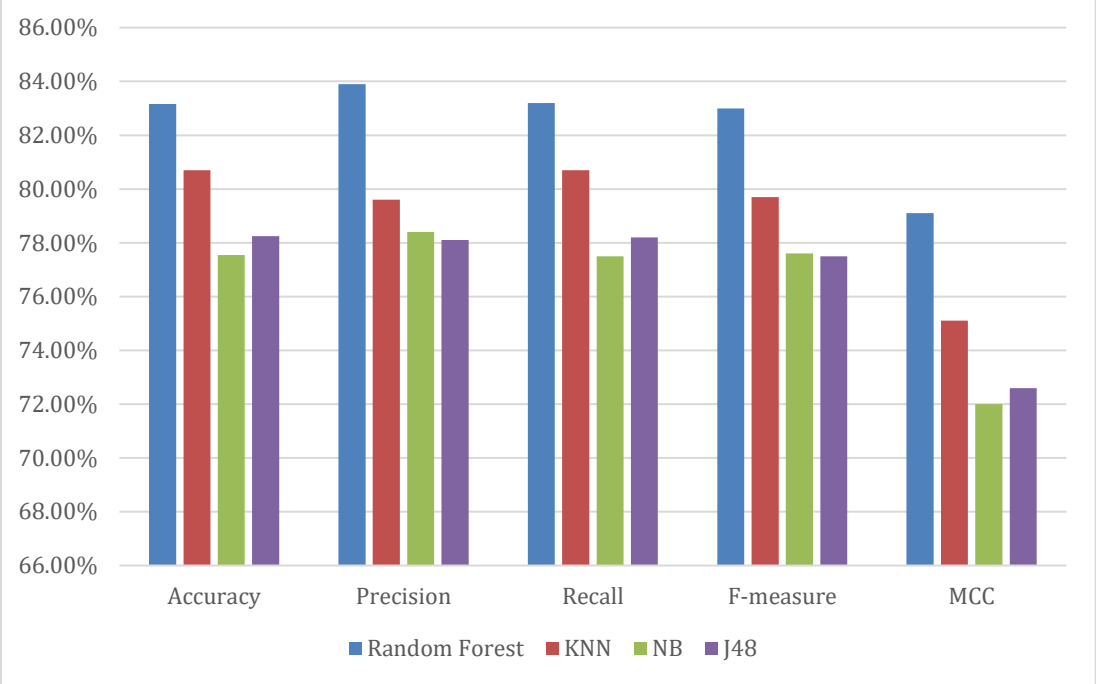

Figure 4: Comparison of Random Forest, KNN, NB and J48 before feature selection.

Figure 5: Comparison of Random Forest, KNN, NB and J48 after feature selection.

4. Discussion

The results obtained in this research showed that better classification models can be obtained both by handling imbalanced dataset properly and through feature selection as shown in Figures 5 and 6. This research also affirms that feature selection helps to alleviate the high dimensionality of dataset in machine learning that mostly leads to over-fitting which is in support of Oladimeji et al. [3]. It was discovered that Random Forest has the highest Accuracy level and thereby correctly predicted the drop out level of students.

Conclusion

Based on the analysis, it can be inferred that JAMB score is a major determinant for predicting student's class of degree due to its ranking in feature selection though it has to be combined with other features such as Age and other O'level result. It is also evident that gender has nothing to do with the prediction class of degree. The results also show that the Random Forest algorithm performs better than the other algorithms in predicting the possibility of students' drop out. The results obtained is promising as it will assist career advisors and educationalists to advice students as appropriate. This research has been able to use real-world data to evaluate the success of students in tertiary institutions with the use of machine learning approach. The future work can be geared towards large dataset of student dropout at both undergraduate and postgraduate level using data mining techniques like Decision Stump, NDTREE and classification to discover relationship among various factors.

References

[1] Olufemi OT, Adediran AA, Oyediran WO, Factors affecting students' academic performance in colleges of education in Southwest, Nigeria, *British Journal of Education*, *6*(10), 2018, 43-56.

[2] Saheed YK, Oladele TO, Akanni AO, Ibrahim WM, Student performance prediction based on data mining classification techniques, *Nigerian Journal of Technology*, *37*(4), 2018, 1087-1091.

[3] Oladimeji OO, Oladimeji A, Oladimeji O, Classification models for likelihood prediction of diabetes at early stage using feature selection, *Applied Computing and Informatics*, 2021. DOI: 10.1108/ACI-01-2021-0022

[4] Odeniyi LA Tunwase AO, Comparative Study of Decision Tree-based models for Predicting Student Academic Achievement in WASSCE, *Proceedings of the 15th iSTEAMS Research Nexus Conference*, Chrisland University, Abeokuta, Nigeria, 16th – 18th April, 2019. pp 247-256 www.isteams.net - DOI Affix - https://doi.org/ 10.22624/AIMS/iSTEAMS-2019/V15N1P23

[5] Yaacob WFW, Nasir SAM, Yaacob WFW, Sobri NM, Supervised data mining approach for predicting student performance. Indonesia, *Journal Electr. Eng. Comput. Sci*, 16(3), 2019,1584-1592.

[6] Bhardwaj BK, Pal S, Data Mining: A prediction for performance improvement using classification, 2012a. arXiv preprint arXiv:1201.3418.

[7] Baradwaj BK, Pal S, Mining educational data to analyze students' performance, 2012b. arXiv preprint arXiv:1201.3417.

[8] Kostopoulos G, Lipitakis AD, Kotsiantis S, Gravvanis G, Predicting student performance in distance higher education using active learning, In *International conference on engineering applications of neural networks*, 2017, (pp. 75-86). Springer, Cham.

[9] Yadav SK, Bharadwaj BK and Pal S, "Mining Educational Data to Predict Student's Retention: A Comparative Study", *International Journal of Computer Science and Information Security (IJCSIS)*, Vol. 10, No. 2, 2012.

[10] Pal S, Mining educational data to reduce dropout rates of engineering students, *International Journal of Information Engineering and Electronic Business*, 4(2), .2012.

[11] Yadav SK, Pal S, Data mining: A prediction for performance improvement of engineering students using classification, 2012. arXiv preprint arXiv:1203.3832

[12] Yadav SK, Bharadwaj BK, Pal S, "Mining Educational Data to Predict Student's Retention: A Comparative Study", *International Journal of Computer Science and Information Security* (IJCSIS), Vol. 10, No. 2, 2012.

[13] Merceron A, Yacef K, Revisiting interestingness of strong symmetric association rules in educational data, In Proc. of Int. *Workshop on Applying Data Mining in e-Learning*, Creete, Greece, 2007, (pp. 3-12).

[14] Navamani JMA, Kannammal A, Predicting performance of schools by applying data mining techniques on public examination results, *Research Journal of Applied Sciences, Engineering and Technology*, 9(4), 2015, 262-271.

[15] Mishra S, Tyagi AK, The Role of Machine Learning Techniques in Internet of Things-Based Cloud Applications, *In Artificial Intelligence-based Internet of Things Systems*, 2022, (pp. 105-135). Springer, Cham

[16] Baashar Y, Alkawsi G, Mustafa A, Alkahtani AA, Alsariera YA, Ali AQ, Tiong SK, Toward Predicting Student's Academic Performance Using Artificial Neural Networks (ANNs), *Journal of Applied Sciences*, 12(3), 2022, 1289.

[17] Adekitan AI, Noma-Osaghae E, Data mining approach to predicting the performance of first year student in a university using the admission requirements, *Education and Information Technologies*, 24(2), 2019, 1527-1543.

[18] Chen W, Xie XS, Wang JL, Pradhan B, Hong HY, Bui DT, Duan Z, Ma JQ, A Comparative Study of Logistic Model tree, Random Forest, and Classification and Regression Tree Models for Spatial Prediction of Landslide Susceptibility. *Catena*, 151, 2017. 147 j.catena.2016.11 .032.

[19] Sweeney M, Rangwala H, Lester J, Johri A, Next-term student performance prediction: A recommender systems approach, 2016. arXiv preprint arXiv:1604.01840.

[20] Hu G, Xi T, Mohammed F, Miao H, Classification of wine quality with imbalanced data. 2016 IEEE *International Conference on Industrial Technology (ICIT*). 2016; 1712-1217. doi:10.1109/ICIT.2016.7475021.

[21] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Bio Bio Acta (BBA)-Pro Struct, 405(2), 1975, 442–51. doi:10.1016/00052797(75)90109-9