ISSN No:-2456-2165

Using Big Data to Determine Potential Dropout of Students in Some Selected Tertiary Institutions in Kebbi State, Nigeria

Bashar Badamasi Lailaba[1]*, Shamsu Sani[2]*, Saifullahi Ahmad Tijjani[3]*, Hassan A[4]*
[1]Department of Sciences Kebbi State Polytechnic, Dakin-gari, Nigeria
[2]Department of Computer Science Kebbi State Polytechnic, Dakin-gri, Nigeria
[3]Department of Computer Science College of Advance Studies Yelwa-Yaur, Nigeria
[4]Department of Mathematics Federal University Birnin Kebbi, Nigeria

Abstract:- Being the essential component of modernity Big data has drawn a lot of interest from practitioners, scholars, and businesses. Given the significance of the education sector, there is a current trend to investigate how big data might be used in this industry to forecast learning results. Student dropout is a significant issue in higher education, affecting both universities and polytechnics. Time to graduation (TTG), which has a direct correlation with student dropout, is one of the key measures of university achievement even if there is no universally accepted way to measure the quality of education (Pineda Lezama, O., & Gómez Dorta, R. 2017). This declining rate indicates a percentage that results in losses of millions to billions of dollars on a global and state level. Yet, as society demands the contributions made by the population with higher education, such as: innovation, knowledge production, and scientific discovery, dropping out has an impact not only on the nation's economy and educational quality but also on the advancement of society. This offers a straightforward method for predicting potential dropouts based on their academic and demographic traits using fundamental statistical learning techniques. The study will be carried out at a few chosen tertiary institutions in Kebbi State.

Keywords:- *Big Data, Demography, Dakin-Gari*

I. INTRODUCTION

In recent times, there has been a sharp rise in the number of students being admitted to postsecondary schools, accompanied by an exponential decline in the number of graduates from these institutions. In higher institutions in Kebbi State, there are multiple surveys that determine dropout rates. The majority of them deal with identifying the causes of dropouts, counting the number of students who do so, and developing strategies to lower the rate (Ahuja, R.; Kankane, Y. 2017). In this work, we present the two schemes that calculate the likelihood that a student will graduate or drop out: the first is based on the percentage of students who graduate within a specific time frame, which corresponds to the time it takes to receive a diploma or degree; the second simply counts the number of students who drop out of their research. These studies suggest strategies for early detection of probable dropout students in an effort to decrease dropout rates. Numerous scholars have created a number of statistical learning methods apps to investigate course completion or dropout rates in order to solve the issue of students leaving school early. These researchers' techniques include, but are not limited to, logistic regression, k-nearest neighbors, decision trees with random forests, Bayesian networks, and neural networks. These Studies, however, are somewhat deficient in clarity and interpretation. The research's suggested methodology will strike a compromise between interpretations and precision. Naïve Bayes and K-nearest neighbors are two methods that give significant precision capacity, and decision trees and logistics regression will be the two methods employed in this work to create the models. When these four approaches are used, a compromise solution between precision and comprehensibility will be produced, with the latter being assessed predominantly by the proportion of dropouts that were found (Bucci, et al., 2018). In this effort, we will leverage the previously mentioned techniques to gather datasets from various state organizations and create an early detection system framework for prospective dropouts. There is little research on student dropouts from Nigerian tertiary institutions, particularly in Kebbi State, despite the abundance of results in the literature regarding the factors that have been shown to impact student dropouts. In order to model student dropout using student admission data from 2016–2022, which was gathered from academic databases, this research presents a number of machine learning algorithms. It also provides comprehensive data about the number of students who dropped out or completed the course, as well as analyses of the causes of dropping out.

II. LITERATURE REVIEW

Schools that have a higher percentage of graduates tend to have a higher number of highly qualified teachers and fewer children from poor backgrounds (Allensworth, 2005; Balfanz, Herzog, & Mac Iver, 2007). Additional factors that also significantly predict high school graduation include being older than average, having subpar grades, and having low attendance. According to Allensworth's research,

in the past, academics have identified the specific characteristics that influence a person's propensity to drop out using analytical techniques like logistic regression. Creating a method to categorize ninth-grade pupils in Chicago's public school system as "on track" or "not on track" for graduation was one of the greatest attempts at this.

A risk measure proposed by Neild and Balfanz (2006) accurately predicted eighth graders' high school graduation in 75% of cases. Despite system flaws, risk prediction analyses have led to the widespread implementation of "early warning" systems in US school districts.

One effective method of reducing dropout rates is to forecast future student dropout rates. In studies on student retention, Tinto's model [2] is most frequently applied. Tinto came to the conclusion that a student's decision to continue or discontinue their studies is significantly influenced by their level of academic and social integration at the institution.

After putting the Tinto model to the test in [10], Brunsden et al. concluded that it might not be the ideal option for dropout study. Durso and Cunha [11] did a study to find the explanatory variables for undergraduate accounting program dropouts at a public institution in Brazil. The survey database used contained socioeconomic and demographic information about 371 pupils. The study's suggested logistic regression model accurately predicted 77% of the sample's incidences of dropout or completion. Five semi-structured interviews with sample members who dropped out of school were used in a qualitative investigation. The findings of the study have improved our comprehension of the issue of undergraduate dropouts from accounting programs and have brought attention to the need for reassessing laws designed to keep talented people in the country, especially those students who work to finance their education.

Kim and Kim [12] conducted a study to examine the possible causes of South Korean university dropout rates. Resources, students, faculty, and university features were the four main areas of concentration. They calculated them using nonlinear panel data models utilizing three-year balanced panel data from 2013 to 2015. The findings demonstrated the considerable effects of teacher quality and quantity, institution size and type, and cost and burden on students' financial resources on university dropout. Numerous other research have also employed data mining techniques to forecast student dropout rates.

Tan and Shao [13] selected the personal characteristics of the students and utilized the Artificial Neural Network (ANN), Decision Tree (DT), and Bayesian Networks (BNs) technique as a prediction model, by choosing the academic achievement and personal traits of the students as input attributions. The outcomes demonstrated that while all three machine learning techniques were successful in predicting student dropout rates, DT performed better.

In the meantime, Mustafa, Chowdhury, and Kamal [14] created a dynamic dropout prediction model for colleges, universities, and institutes using data mining. Gender, financial status, and year of dropout were utilized as classification factors to separate successful from unsuccessful students. The factors were examined using Classification and Regression Tree (CART) and CHAID following data separation. With the highest percentage of accurate classification overall, CART outperformed CHAID in tree growth.

Yukselturk, Ozekes, and Türel [9] looked studied the data mining techniques used in an online application to forecast dropout rates in a different study. Gender, age, education level, prior online experience, occupation, self-efficacy, readiness, prior knowledge, locus of control, and dropout status were the variables included in this study. Four data mining techniques—k-Nearest Neighbor (k-NN), Decision Tree (DT), Naïve Bayes (NB), and Neural Network (NN)—were used to categorize students who had dropped out. Every approach was trained and tested using 10-fold cross validation. The 3-NN, DT, NN, and NB classifiers have detection sensitivities of 87%, 79.7%, 76.8%, and 73.9%, in that order. The most important variables were found to be self-efficacy, readiness for online learning, and previous online experience using the Generic Algorithm (GA), The most important criteria in predicting dropout rates were found to be self-efficacy, readiness for online learning, and prior online experience.

A total of 1290 computer science graduates from ALAQSA University between 2005 and 2011 were examined by Abu-Oda and El-Halees [15] using various data mining techniques to examine and predict students' dropout rates. The data sets were subjected to the application of various classifiers, including Decision Tree and Naive Bayes, and they were tested using 10-fold cross validation. 98.14% and 96.89%, respectively, of the classifiers were accurate. The underlying links between students' dropout status and persistence in their enrollment were also discovered using the FP-growth algorithm. The findings demonstrated a strong correlation between learning "digital design" and "algorithm analysis" courses and student success.

300 undergraduate computer course participants from five different universities were evaluated for their performance using EDM by Bharadwaj and Pal [4]. The results of a senior secondary test, place of residence, different habits, annual family income, and family status were found to be significant predictors of academic achievement in their Bayesian categorization system of 17 variables.

In order to predict academic achievement, Bharadwaj and Pal [5] created a new data set in a follow-up analysis that includes test, seminar, and assignment marks in addition to student attendance. Kovacic [6] suggested a study akin to this, using EDM to determine which enrollment information may be utilized to forecast students' academic achievement.

ISSN No:-2456-2165

He applied the CART and CHAID algorithms on a dataset of a student enrolment.

In an attempt to improve the caliber of higher education, Al-Radaideh et al. [16] evaluated student academic data (student gender, student age, department, high school grade, lecturer degree, lecturer gender, among others) using a classification model constructed using the decision tree approach. They found that the feature with the largest gain ratio, the high school graduation rate, was considered the root node of the decision tree. The Holdout method and the K-Cross-Validation method (k-CV) were used to evaluate the model. But they found that the collected samples and attributes were not enough to generate a high-quality classification model.

In a case study, Gerben et al. [17] projected student accomplishment using machine learning techniques and characteristics extracted from students' pre-university academic records. Decision trees are an easy-to-understand and intuitive classifier that yielded practical results with accuracy levels ranging from 75% to 80%, according to their testing findings. One of their findings was that, even though it wasn't thought of as the essential course, linear algebra was the best indicator of success. Despite these findings, it's not apparent which data mining techniques work best in this scenario.

Luan, for example, developed predictors in [18] by using clustering as a method for data exploration and classification. One of the findings of Romero and Ventura's survey on EDM [19] was that association analysis has become a popular tactic.

Aulck et al. [20] used the largest available database of higher education attrition to estimate student dropout using transcript data and demographic data from 32500 students at one of the major public universities in the nation. According to the results, it is possible to accurately forecast several early indicators of student attrition and dropout, even when assumptions are based on data from academic transcripts spanning only one term. It raised awareness of the implications of student retention and success and retention using artificial intelligence.

III. METHODOLOGY

This section presents the process of data collection and analysis of the collected data.

➢ *Data Collection*

In this section data were collected from the admission portal of the selected schools which includes the demographic information (Gender, Ag, support and resources, occupation, and challenges and Barrier) and examination records from the selected school. The examination records were collected with authorized permission from Examination Department of those Institutions. The two records were joined together for data cleansing and analysis.

The data that was collected from the selected schools, there were 99,867 records in all, equivalent to the enrolments of 20,807 students. Records that were incomplete or contained incorrect information were deleted after undergoing the data cleaning, resulting in a final sample of 85,527 records of 17,720 students in all. In other to be able to predict the dropouts, four assumptions were used, each of which involved the analysis of different numbers of records:

- The first assumption uses data from all records for students who enrolled in the semesters between the years 2016 to 2022. A dropout is assumed to be a student who is yet to graduate and spent at least two years without enrolling into the program, while a non-dropout is an active student or one who finished his or her studies between 2016 and 2022. A total of 85,527 records satisfied these criteria, of which 28.1% belonged to students who were classified as dropouts.
- In the second assumption a dropout is defined in the same way as in the previous perspective; however, a non-dropout is defined as a student who finished his or her studies. The objective here is to eliminate noise of active students when training the algorithm, since it is not known beforehand if they will graduate or abandon. A total of 35,132 records satisfied these criteria, of which 43.7% belonged to students who were defined as dropouts.
- In the third assumption, dropouts and non-dropouts are defined as in perspective 1. The difference is that only one semester (one period) is used for each student who enrolled between 2016 and 2022. In the case of dropouts, only information of the last semester before dropping out is used, and in the case of non-dropouts, a semester is chosen at random. The purpose of this perspective is to eliminate noise from previous semesters of the dropout, on the assumption that the most recent semester provides the most up-to-date information to predict if he or she is going to drop out. A total of 15,720 records satisfied these criteria, of which 28% belong to students who were defined as dropouts.
- In this assumption, the definition of dropouts and non-dropouts is the same as that used in perspective 2, but only one semester per student is used, as was done in third perspective. The objective of this perspective is to eliminate noise from active students and from previous semesters of students who drop out. A total of 7,936 records satisfied these criteria, of which 55.7% belong to students who were defined as dropouts.

Table 1 presents the percentages of all participants' demographic characteristics. The number of male students (70%) was greater than the number of female (30%) students, and the students' ages ranged from 20 to 50 with an average of 24. The majority of the students were undergraduate and graduate student (Diploma or HND students) (60.3%). Nearly half of the students (49.7%) have full-time or part-time jobs and only a few of them (10.5%) have sponsorship support.

ISSN No:-2456-2165

Table 1: The Demographic Characteristics of Participants

Gender	Number of registered participants	Number of dropout participants	percentage of registered participants	percentage of dropout participants
First Assumption				
Female	25658.1	2163.00	30	8.43
Male	59868.9	3309.08	70	19.67
Second Assumption				
Female	9074.60	1189.68	25.83	13.11
Male	26057.40	7971.00	74.17	30.59
Third Assumption				
Female	2751	231.10	17.50	8.4
Male	12969	2541.92	82.50	19.6
Fourth Assumption				
Female	3902.13	652.05	49.17	16.71
Male	4033.87	1572.81	50.83	38.99
Age				
20-39	70,559.78	55,780.71	82.50	65.22
30-50	10,690.88	23,554.12	12.50	27.54
Sponsorship				
YES	8980.335			
NO	76,546.67			

IV. DISCUSSION OF THE RESULTS

In this section, results obtained from the analysis are presented to show the percentage and the distribution of student's dropout.

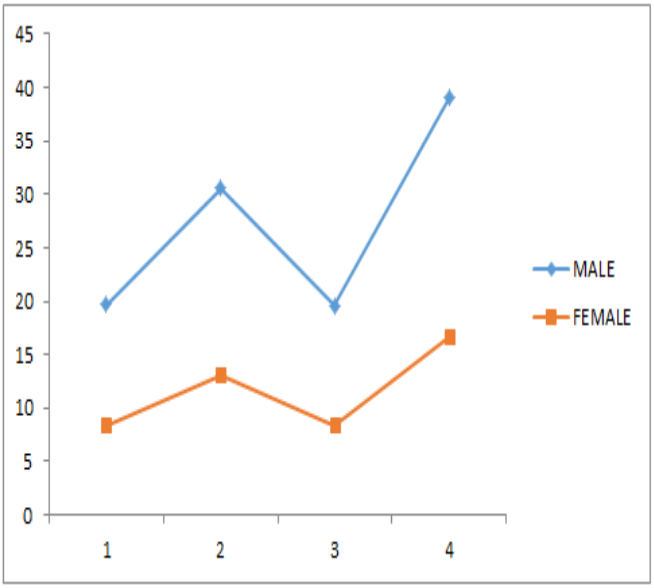

Fig 1: percentage Dropout for Male and Female based on the four assumptions

Figure 1, shows the trend in dropout rate of students who got admitted for the period 2016-2022, classified by gender based on the above four assumption. The dropout rate for male is high compare to the dropout rate of Female; however, the dropout rates of female are still high percentages specially if translated into absolute amounts also it is because the number of female admitted is always less than that of male.

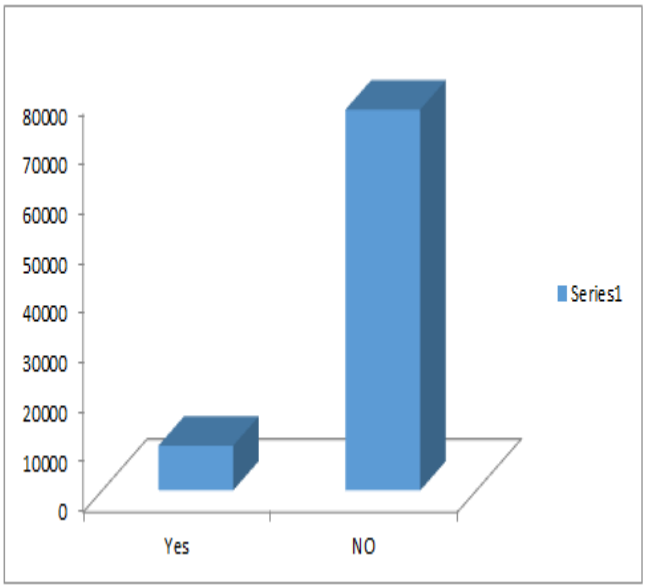

Fig 2: Distribution of scholarship student against Non-scholarship student

Figure 2. Shows the distribution of scholarship students and no-scholarship students from the distribution one can clearly see that non-scholarship students are higher than those with scholarship this is one of the contributing factor that leads to high number of dropout around the

ISSN No:-2456-2165

schools due to economical factors that affects the leaving standard of those students.

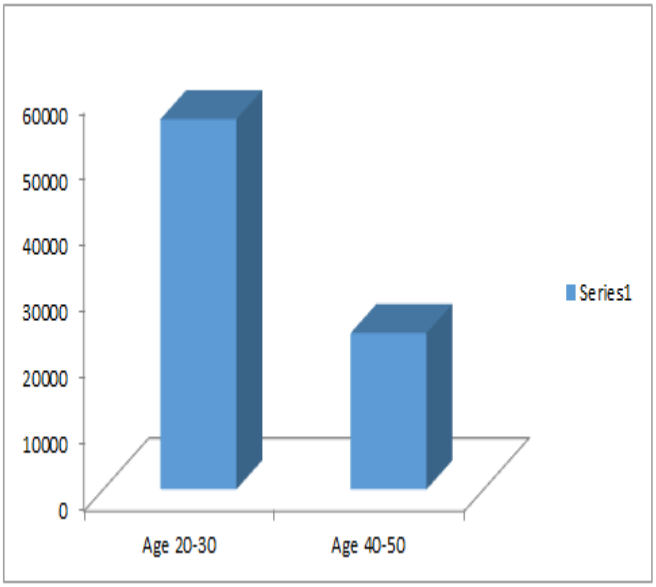

Fig 3. Showing the Distribution of dropout by age

From the distribution above it can be seen that the number of students between the ages of 20-30 has high number of dropout this as a result of the fact that most of them lacks academic, financial, and emotional support to make them persevere in their studies. While the student between the ages of 40-50 has low number of dropout compare to the former one also here, the dropout is as a result of challenges they faced at their working place, family challenges and health related issues due to their ages.

V. CONCLUSIONS

The high school dropout rate among students is a key performance indicator for higher institutions. Telling the student that they are in danger, are more likely to study more in tensely and better organize their semester workload. The level coordinator can use the accurate prediction to determine whether or not to approve or deny students' requests to repeat failed courses.

ACKNOWLEGEMENT

This work was funded by a grant for Institutionbased research (IBR) 2022 and 2023 marger for Kebbi State Polytechnic Dakingari from the Nigeria's Tertiary Education Trust Fund (Tetfund).

REFERECES

[1]. Tinto, V.: Dropout from higher education: a theoretical synthesis of recent research. Rev. Educ. Res. 45(1), 89-125 (1975) 22. Wirth, R.: CRISP-DM: towards a standard process model for data mining. In: Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, pp. 29-39 (2000)

[2]. Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Mining Education data to predict student's retention: a comparative study. arXiv preprint arXiv:1203.2987.

[3]. Baradwaj, B.K., Pal, S.: Mining educational data to analyze students' performance. arXiv preprint arXiv:1201.3417 (2012)

[4]. Bhardwaj, B.K., Pal, S. (2012) Data mining: a prediction for performance improvement using classification. arXiv preprint arXiv:1201.3418

[5]. Kovacic, Z. (2010). Early prediction of student success: Mining students' enrolment data.

[6]. Devasia, T., Vinushree, T. P., &Hegde, V. (2016, March). Prediction of students performance using Educational Data Mining. In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE) (pp. 91-95). IEEE

[7]. Tekin, A. (2014). Early prediction of students' grade point averages at graduation: A data mining approach. Eurasian Journal of Educational Research, 54, 207-226.

[8]. Yukselturk, E., Ozekes, S., &Türel, Y. K. (2014). Predicting dropout student: an application of data mining methods in an online education program. European Journal of Open, Distance and e-learning, 17(1), 118-133.

[9]. Brunsden, V., Davies, M., Shevlin, M., & Bracken, M. (2000). Why do HE students drop out? A test of Tinto's model. Journal of further and Higher Education, 24(3), 301-310.

[10]. Durso, S. D. O., & Cunha, J. V. A. D. (2018). Determinant factors for undergraduate student's dropout in an accounting studies department of a Brazilian public university. EducaçãoemRevista, 34.

[11]. Kim, D., & Kim, S. (2018). Sustainable education: Analyzing the determinants of university student dropout by nonlinear panel data models. Sustainability, 10(4), 954.

[12]. Tan, M. & Shao P. (2015). Prediction of student dropout in e-learning program through the use of machine learning method. International Journal of Emerging Technologies in Learning (iJET), 10(1), 11-17.

[13]. Mustafa M. N., Chowdhury L., & Kamal M. S. (2012, May). Students dropout prediction for intelligent system from tertiary level in developing country. In 2012 International Conference on Informatics, Electronics & Vision (ICIEV) (pp. 113-118). IEEE.

[14]. Abu-Oda G. S. & El-Halees A. M. (2015). Data mining in higher education : University student dropout case study. International Journal of Data Mining & Knowledge Management Process (IJDKP), 10(1), 15-27.
[15]. Al-Radaideh, Q. A., Al-Shawakfa, E. M., & Al-Najjar, M. I. (2006, December). Mining student data using decision trees. In International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan.
[16]. Dekker, G. W., Pechenizkiy, M., &Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. International Working Group on Educational Data Mining.
[17]. Jing, L.: Data mining and its applications in higher education. New Dir. Inst. Res. 2002(113), 1736 (2002). https://doi.org/10.1002/ir.35, https://onlinelibrary. wiley.com/doi/abs/10.1002/ir.35
[18]. Romero, C., Ventura, S.: Educational data mining: a survey from 1995 to 2005. Expert Syst. Appl. 33(1), 135 - 146 (2007). https://doi.org/10.1016/j.eswa.2006.04.005,http://www.sciencedirect.com/science/article/pii/S0957417406001266 Applying DM Techniques to Predict Student Dropout 125
[19]. Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. arXiv preprint arXiv:1606.06364.
[20]. Herzog, S. (2006). Estimating student retention and degree?completion time: Decision trees and neural networks vis?à?vis regression. New directions for institutional research, 2006(131), 1733.
[21]. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS inc, 16.Seidman, A.: Retention revisited: R= E, Id+ E & In, Iv. Coll. Univ. 71(4), 18-20 (1996)
[22]. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.