FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 7 No. 3, June (Special Issue), 2023, pp 177 - 185
DOI: https://doi.org/10.33003/fjs-2023-0703-1773

PREDICTING TIMELY GRADUATION OF POSTGRADUATE STUDENTS USING RANDOM FORESTS ENSEMBLE METHOD

[1]Hafsat S. Bako, [*2]Faruku U. Ambursa, [3]Bashir S. Galadanci, [4]Muhammad Garba

[1]Computer Science Department, Bayero University Kano, Nigeria
[2]Information Technology Department, Bayero University Kano, Nigeria
[3]Software Engineering Department, Bayero University Kano, Nigeria
[4]Computer Science Department, Federal University Brinin Kebbi, Kebbi, Nigeria

*Corresponding authors' email: fuambursa.it@buk.edu.ng

ABSTRACT

Graduation time of students, both undergraduate and postgraduate, has been a prime focus in universities recently. Over the years, there have been numerous research on using data mining techniques to forecast undergrad students' success. However, very few works have been reported on predicting graduation time of postgrads, particularly using data from Nigerian Universities. This research utilized classification techniques using supervised learning to develop a Postgraduate Student Graduation Time Prediction Model (PS_GTPM). Data was collected from Bayero University Kano and the Adaptive synthetic sampling (ADASYN) technique was applied to address the imbalance issue with the data. Then, the model was developed using the Random Forests ensemble technique. From the evaluation results, we found that the data balancing method based on ADASYN technique enhanced the ability of the data mining classifiers to forecast when students will graduate. Also, it was found that the proposed PS_GTPM based on Random Forests Ensemble Method recorded the highest prediction accuracy with more than 83% score compared to the other methods. Largely, PS_GTPM can be used to forecast whether a thesis-based graduate study shall be completed on-time or not.

Keywords: Educational Data Mining (EDM), Student Performance Prediction, Machine Learning, Ensemble Learning

INTRODUCTION

Postgraduate students' graduation time is an important concern for universities, locally in Nigeria and around the world. This is due to the fact that a high graduation rate is one of the most important indicators for ranking universities in the education sector just as it also reflects significantly in its annual operation costs (Nurafifah et al., 2019) (Suhaimi et al., 2019) . In Nigeria, although the Federal Ministry of Education, through the Universities, is implementing diverse strategies and initiatives to ensure postgraduates complete their programme on time, the situation is still worrisome. Although the factors responsible for this predicament vary and sometimes delay graduation is unavoidable, careful study of the situation would yield beneficial outcomes (Suhaimi et al., 2019). Fortunately, numerous pieces of information regarding the traits and performance of students are gathered by universities and other institutions over the years, and useful information can be extracted from these historical data to gain insights and predict future data. While there have been numerous works on predicting the graduation of postgraduate research students in different countries, very little research has been conducted on data from Nigerian Universities. Models developed for use in other countries may not apply well in Nigeria because of the peculiarities in terms of our educational system as well as the factors affecting the students' studies. Hence, the purpose of this study, is to propose a model for predicting graduation time for postgraduate students' in Nigerian universities using historical records.

To develop prediction models, Education Data Mining (EDM) comes into play as it can be used in discovering patterns for making decisions from educational data (Osmanbegovic & Suljic, 2012). Prediction, relationship mining, clustering, data distillation for human judgment, and model-based discovery are classes of EDM (Baker & Yacef, 2009). The goal of prediction is to make forecast about unknown variables using historical data for the same variable type. The types of patterns to be found in the data mining process are specified using data mining techniques such as class description, association analysis, classification and prediction, cluster analysis, and outlier analysis. Finding new models to explain and differentiate data classes and concepts in order to predict unknown classes of objects is the process of classification. Data items' class label can be predicted via classification. To perform prediction, numerous single classifier algorithms exist such as the Decision Tree (DT), k-Nearest Neighbor (k-NN), sequential minimal optimization (SMO), Logistic Regression (LogR), Naïve Bayes (NB), Support Vector Machine (SVM) etc. In this research, however, these approaches based on single classifiers yield less accurate results (Finlay, 2011). A better and trending approach is the use of ensemble learning. Ensemble learning is constructing a prediction model by juxtaposing the strengths of a number of other prediction models.

In this research a Postgraduate Student Graduation Time Prediction Model (PS_GTPM) is proposed. To develop the model, first, data was collected from Bayero University, Kano, a large public university in Nigeria's second largest city with a student population of over 45,000 out of which close to 11,000 are postgraduate students. Features including name, registration number, course of study, phone number, cumulative grade point average (CGPA), age, gender, nationality, state of origin, home address, sponsor and marital status, year of entry, and year of graduation were collected. We used the WEKA filter to remove all unwanted entries from the dataset. All information related to students' personal identity was removed. The data was cleaned to get only the relevant features for this work which reduced the data attributes to age, gender, marital status, CGPA, year of entry, and year of graduation. Moreover, the data we collected had

class imbalance problems, where the class distribution of the dataset is imbalanced. Class imbalance problems significantly impair the performance of standard classifiers (Johnson and Khoshgoftaar, 2019). To deal with the imbalanced data we used Adaptive synthetic sampling (ADASYN) technique. The goal of ADASYN is to insert a higher ratio of synthetic data close to the minority class points that a model would find most challenging to learn. To build the prediction model, we employ the Decision Tree algorithm. The tree-based method is chosen for its efficiency in dealing with categorical data such as the type collected for this work (Gareth et al., 2015). Also, the decision tree is fast, flexible and supports feature interaction better than most other methods. However, simple decision trees tend to overfit the training data and the tree may grow to be very complex while training complicated datasets. To address these problems, ensemble learning, specifically the Random Forests ensemble method, was used. The Random Forests classification was chosen for its inherent efficiency and capability in handling statistical noise. Majority voting technique was used for the classification. For model training and testing, 10-fold cross validation was used. In order to show the effectiveness of our proposed PS_GTPM, we compare the model with other popular ensemble classifiers such as bagging, boosting and random tree ensembles. The performances of the various models were evaluated based on a confusion matrix using accuracy, precision, recall and f-measure metrics. From the results of experiments we found that the data balancing method based on ADASYN technique improved the performance of the data mining classifiers in predicting students' graduation time. It was also discovered that ensemble learning performed better than the individual classifiers on decision trees. Furthermore, of the four ensemble models evaluated, it was found that the proposed Random Forests based PS_GTPM recorded the highest prediction accuracy of more than 83%.

Related Works

This section gives a review of works done on predicting students' academic performance, especially their graduation times. This has been one of the several problems in EDM and classification has been used the most in solving these kinds of problems (Thakar, 2015). In a study to predict graduation rates of postgraduate students Goenner and Snaith (2004) studied how institutional factors affect doctorate universities' graduation rates. Using multivariate regression analysis, they proposed a model to investigate the factors that influence overall graduation rates. (Agu and Oluwatayo, 2013) conducted quantitative analysis to determine why many postgraduate distance learners don't finish their dissertations after taking the required courses. To get perspectives from some postgraduate distance learners about the factors influencing their completion of research work. They used structured questionnaires created on a five-point Likert-type scale. The study came to a conclusion with some suggestions on how to improve the administration of research work writing by distance learners. The reference (Tampakas et al., 2014) conducted research to identify the factors linked to postgraduate students' delayed thesis completion, which results in a prolongation of the graduation period. The research used a descriptive survey approach. Students were chosen through snowball sampling, and the researchers modified and validated the POSTDAQ questionnaire. The study demonstrates that the factors associated to students are more responsible for the delay in finishing the thesis. Gbolagade et al., (2015) used two level classification algorithms to predict students' graduation time. The proposed algorithm first determines those students who are most at danger of not finishing their studies, and then groups students based on their anticipated graduation dates. Shariff et al. (2016) employed the sequential minimum optimization method (SMO), radial basis function (RBF), and multilayer perceptron (MLP) to categorize data and portrayed the impact of data pre-processing algorithms. To evaluate the factors that determine possible PhD candidates, attribute selection was used. The study reveals that the performances of MSc. students are determined by State of origin, marital status, higher qualifications and gender of the student. Hadi and Muhammad (2019) proposed a model to predict the number of Ph.D students that will complete their study on time. The study employed binary logistic regression on a set of data. Their result showed that only 6.8% of the 2014 Ph.D candidates were predicted to graduate on time. Suhaimi et al. (2019) built a prediction model that forecast students' chances of graduating. The authors used the Support Vector Machine (RBFKernel), Support Vector Machine (PolyKernel), Decision Tree, Random Forest, and Naive Bayes machine learning algorithms, which were applied separately. The classifiers' performance in terms of accuracy, precision, recall, and F-Score performances were assessed and compared. Support Vector Machine (PolyKernel) fared better than other classifiers, according to the results. However, the study was based on undergraduate data. Ahmad et al. (2015) determined the factors influencing performance of graduate research students. They used exploratory factor analysis to identify and statistically analyze 41 indicators and 112 valid responses. Higher-order factors were discovered and statistically assessed using variance-based structural equation modeling. The research discovered positive and significant correlations between the performance of research students and institutional, student, and supervisor-related factors. Also, the findings showed that students' personal factors had significant impact on performance of research students. This is followed by institutional factors and then supervisor-related factors. Numerous other researches (Knutson, 2020; Amida et al., 2020; Verostek et al., 2021; Agbonlahor, 2022; Muthukrishnan et al., 2022) have been conducted based on current data and statistical methods. However, our work differs in that it is based on historical data rather than exploratory data.

Recently, Baashar et al. (2022) proposed a model that predicted cumulative grade point average (CGPA) of postgraduate students at masters level using machine learning (ML). They collected real historical dataset of 635 master's students from a private university in Malaysia. They applied six ML models and found that Artificial Neural Network (ANN) recorded best performance.

While there have been numerous works on predicting the timely graduation of students in different countries, very few researches have been conducted on the graduation of postgraduate students using historical data here in Nigeria. Models developed for use in other countries may not apply well in Nigeria because of the peculiarities in terms of our educational system as well as the factors affecting the students' studies. From the highlighted problems, it is evident that predicting postgraduate students' graduation time using data from Nigerian universities is required and more accurate prediction models need to be developed for making such predictions.

MATERIALS AND METHODS

The data used in this research were collected from Bayero University, Kano, a large public University in Nigeria's second largest city with a student population of over 45,000 out of which close to 11,000 are postgraduate students. The

processes of data collection, data preparation, feature selection, data balancing, ensemble methods and how the evaluation experiment was carried out are explained.

Data Collection

Data for academic masters' students (M.Sc., MA) were collected from the different faculties of the University. A total of 972 students' records were obtained from the faculties of Science, Engineering, Medicine, Education, Arts and Islamic Studies, Computer Science and Information Technology, Agriculture and Law covering the academic sessions from 2011/2012 to 2017/2018. A wide range of information was collected including name, registration number, course of study, phone number, cgpa, age, gender, nationality, state of origin, home address, sponsor, marital status, year of entry and year of graduation. The study only considered semesters that the students were engaged in school. Therefore, deferred semesters were not taken into consideration. Other factors, including socio-economic and psychometric issues such as strike and personality traits were not included in this research. It is also important to note that the model was designed to predict students' graduation time, not their grades or academic performance. Table 1 presents the detailed description of the collected dataset.

Table 1: Data description

Feature	Value	# Record	Type	Description
Age	25-40 years	679	N*	Students age
	41-69 years	293		
Gender	Male	689	Nm**	The gender of the students
	Female	283		
Marital Status	Married	695	Nm	The marital status of the students
	Single	274		
	Divorced	2		
	Widowed	1		
Duration of Study	2-3 years	392	N	The time it took the students to graduate
	4-6 years	580		
CGPA	2.04-3.49	176	N	The cumulative grade point average of the students
	3.5-5.00	796		
Class	On-time	392	Nm	The class label
	Not-on-time	580		

*N = Numeric **NM = Nominal

Bayero University Students record

Data Preparation

This stage involves preparing the data for analysis. Data collected was initially written in MS Excel sheets. Data was first converted into CSV format for it to be accepted by the Waikato environment for knowledge analysis (WEKA) explorer. Data mining requires clean, preprocessed data given that real-world data is frequently messy, inconsistent, and noisy. We used the attribute WEKA filter to remove all unwanted entries from our dataset. All information related to students' personal identity was removed. The data was cleaned to get only the relevant features for this work. The preprocessing of the dataset was done at this stage before the data mining techniques were used.

Data mining often requires merging data from different sources. In this study, we merged the data we collected from the database with the ones we collected manually. Both records were merged and written in Microsoft Excel. The data was then transformed into forms appropriate for mining. As WEKA only accepts .arff and .csv files, we converted our records from an Excel worksheet having an .xlsx extension to a .csv (comma separated values) file, which was saved in WEKA with an .arff extension.

Feature Selection

Feature selection speeds up the training of machine learning algorithms, simplifies and decreases the complexity of models, increases their accuracy, and lessens overfitting. The feature selection is independent of any machine learning algorithms. Feature selection removes some attributes from the training data while leaving the relevant and useful attributes (Ahmad et al., 2015). We used the wrapper method of feature selection. Name, registration number, course of study, phone number, cumulative grade point average (CGPA), age, gender, nationality, state of origin, home address, sponsor and marital status were collected but after feature selection, the features used in this study were reduced to age, gender, marital status, CGPA, year of entry, and year of graduation.

Data Balancing

Data imbalance is a major issue when utilizing machine learning across several fields (Brennan, 2018). When trained on uneven datasets, machine learning algorithms are prone to producing unreliable predictions. This is a result of the classifier frequently learning to constantly anticipate the class with the majority. Previous researches used SMOTE to balance their datasets (Unal et al., 2019). However, the possibility that surrounding samples may belong to different classes is not taken into account by SMOTE. As a result, there may be more class overlap and add to the noise. SMOTE is also not particularly applicable for data with high dimensions. In this research, to deal with imbalanced data in our preprocessing we used Adaptive synthetic sampling (ADASYN), which is an improved version of SMOTE. The goal of ADASYN is to insert a higher ratio of synthetic data close to the minority class points that would be the most challenging for a model to learn. It makes the sample more realistic by adding a modest random value to the points after it has been created. The fundamental concept of the ADASYN method is to automatically determine the number of synthetic samples that need to be generated for each minority data example by using a density distribution as a criterion. The resulting dataset after ADASYN will compel the learning algorithm to concentrate on those tough to learn cases in addition to offering a balanced representation of the data spread. This is a significant contrast compared to SMOTE approach, which generates an equal number of synthetic samples for each minority data example.

Modeling

In this stage, the prediction model for the graduation time of postgrads is designed. For designing the model, this research uses a machine learning approach, specifically the classification technique, to find patterns in the dataset collected. In order to be able to forecast unknown classes of things, classification is the process of discovering a new set of models that characterizes and separates data classes and concepts. The class label of data objects can be predicted via classification.

To perform prediction, numerous single classifier algorithms exist such as the Naive Bayes (NB), Logistic Regression (LogR), Sequential Minimal Optimization (SMO), Decision Tree (DT), k-Nearest Neighbor (k-NN), Support Vector Machine (SVM), etc. In this paper, we employ the Decision Tree algorithm to build the prediction model. The tree-based method is chosen for its efficiency in dealing with categorical data such as the type collected for this work [8]. Also, the decision tree is fast, flexible and supports feature interaction better than most other methods. Decision trees have an intriguing property in that they can handle qualitative predictions without the need for dummy variables. They just need a data table, and they can create a classifier directly from that data without having to do any up front design work. However, simple decision trees tend to overfit the training data and the tree may grow to be very complex while training complicated datasets. To address these problems, ensemble learning is used. By combining what is good of a number of simpler base models, ensemble learning aims to create a prediction model. Hence, the predictive performance of trees can be greatly enhanced by aggregating multiple decision trees, using ensemble methods. To achieve this, we employ the Random Forests ensemble method, which, using bootstrapped training samples, creates a number of decision tree forests. Every time a split in a decision tree is taken into account, a random sample of m predictors from the entire collection of p predictors are selected as split candidates. Random forests is a variation of the bagging algorithm. It is so named because decision trees are used to build it. (Breiman, 2004). To create a random forest, individual decision trees with specific training settings that fluctuate randomly are required. (Polikar, 2006). The Random Forests was chosen for its inherent efficiency and capability in handling statistical noise.

Majority voting technique was used for the Random Forests classification. When there is a need to classify an input instance, each member (tree) of the ensemble votes for one class label, and the final class label is the one that receives the majority of the votes (half of the votes +1).

Evaluation Experiments

Model Training and Testing

On the dataset of the 972 students, a number of training and testing exercises were performed. To prevent the model from being overfitted, 10-fold cross validation was utilized during model training and testing. The complete dataset is split into ten mutually exclusive sets, each of which has an equal number of students who graduated and those who did not, using a stratified random sample technique. While the data is being processed through the final dataset, nine out of the ten sets were used as training data to create models. For the first model, a classification error rate was computed and saved as an independent test error rate. Next, the test error rate for a second model that was built using a different set of nine samples is calculated. Ten trials of the same procedure resulted in ten different models. After then, the classification error rates for each of the ten models were averaged. To reduce the inaccuracy, the ideal design parameters were selected. Because the mean is more accurate for cross-validation than for a single experiment, accuracy is improved (Nisbet et al., 2009).

In order to show the effectiveness of our approach, we compare it with other popular ensemble classifiers such as bagging, boosting and random tree ensembles. These ensemble methods are briefly described below:

i. Bootstrap Aggregating is referred to as "bagging." In the traditional bagging technique, replacement is used to create 'n' unique bootstrap training samples. After that, the algorithm is trained on each bootstrapped algorithm separately, and then the predictions are aggregated.
ii. AdaBoost: According to Polikar (2006), AdaBoost creates hypotheses by training a weak classifier, then combines them based on weighted majority voting of the classes each hypothesis predicts.
iii. Random tree: The ensemble learning technique known as Random Tree produces numerous individual learners. It creates a decision tree using a bagging idea to generate a random set of data (Kalmegh, 2015).

For this comparison, we chose decision trees, support vector machines and the Bayesian network as base classifiers for the ensembles. The Decision Tree (DT) was chosen given its efficiency and sensitivity to changes in data inputs. On the other hand Sequential Minimal Optimization (SMO) was selected to represent support vector machines because it is fast and Naïve Bayes was chosen to represent the Bayesian networks because of their robustness and simplicity. Each of the compared ensembles trains each of the three classifiers. Majority voting is also adopted for each of the compared models in order to maintain fairness in the comparison.

Evaluation Metrics

A confusion matrix was used to assess how well each model performed. The terms "true positives" (TP) and "false positives" (FP) refer to the proportion of expected positives that are, respectively, true and false. The number of predicted negatives that are correct and incorrect, respectively, is shown by the True and False Negatives (TN and FN, respectively). As a result, the metrics for average accuracy, precision, recall, and f-measure were employed. These measurements are described as follows:

Average Accuracy (Acc): This is the average per-class of a classifier. The accuracy measure shows how well a classifier can anticipate both positive and negative events. It is defined as the average of all the correct number of predictions made (TP +TN) divided by the total number of predictions made, as expressed in equation (1).

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

Precision (Pr): When classifier labels and data class labels are compared on a per-class basis, the average per-class agreement is determined. The following formula measures the ratio between the number of accurately predicted positives and all predicted positives:

$$Pr = \frac{TP}{TP + FP} \quad (2)$$

Recall (Sensitivity R): the data class's average per-class agreement with the class members as determined by a classifier. The proportion of actual positives that were accurately forecasted as positives. It shows how the classifier can find every desired data point in a dataset. (Suhaimi et al., 2019).

$$R = \frac{TP}{TP + FN} \quad (3)$$

F-measure (F_β): the relationships between a classifier's positive labels—those assigned to the data—and those determined by sums of per-class judgments. It is described as the precision and recall's harmonic mean. A classifier that has an F-measure of 100% has, in the best case scenario, precisely balanced precision and recall. The worst case situation, represented by an F-measure of 0, is when the classifier performed poorly in both recall and precision.

$$F_\beta = \frac{(\beta^2 + 1)\, Pr.\, R}{\beta^2\, Pr + R} \quad (0 \leq \beta \leq \infty) \quad (4)$$

Where β is a parameter that controls a balance between Pr and R (Sasaki, 2007). $\beta = 1$, means F-measure has achieved harmonic mean of Pr and R. In contrast, a case of $\beta > 1$ or $\beta < 1$ signifies respectively that F-score is more recall-oriented or precision-oriented.

RESULTS AND DISCUSSION

Results

This section shows the results obtained after conducting the various experiments. Initially we tried the base classifiers with the imbalanced data set and then with ADASYN and compared the results. Table 2 presents the preliminary results of the base classifiers without balancing the data (A1) and after balancing the data (A2) respectively.

Table 2: Preliminary results of base classifiers on imbalanced dataset (A1) and balancing with ADASYN (A2)

Classifiers	**NB**		**SMO**		**J48**	
	A1	A2	A1	A2	A1	A2
Accuracy (%)	56.1	62.4	56.6	60.0	57.1	66.8
Recall (%)	65.6	63.7	68.7	59.9	66.3	67.0
Precision (%)	55.1	85.2	53.3	99.9	56.2	87.2
F-measure (%)	59.9	72.9	61.3	74.9	60.8	75.8

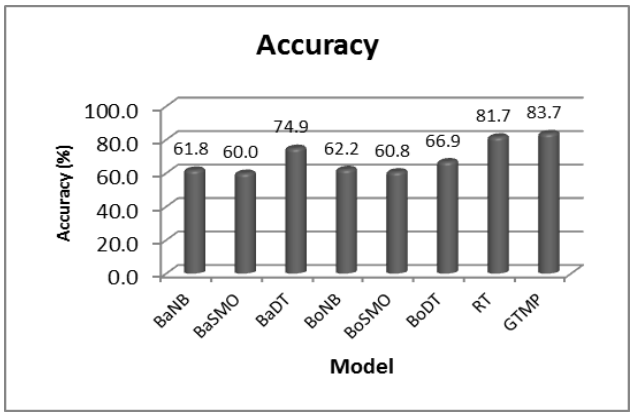

Figure 1: Accuracy score of the different models

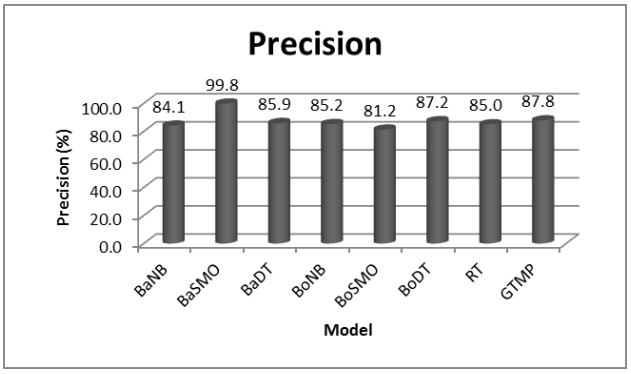

Figure 2: Precision of the different models

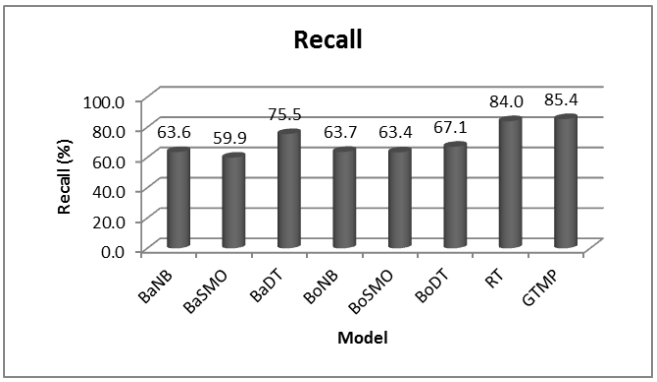

Figure 3: Recall of the different models

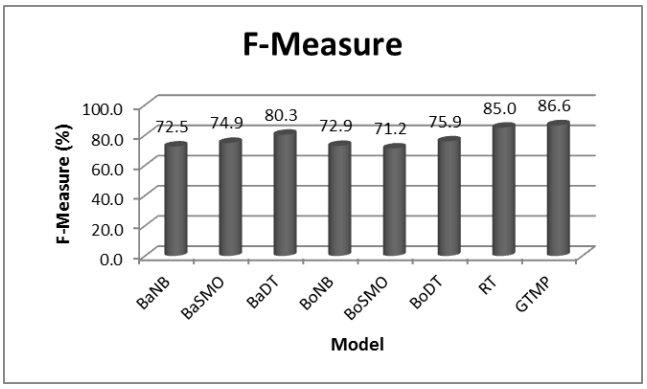

Figure 4: F-Measure of the different models

From Table 2 it is shown that the classifiers that were trained on the dataset that was balanced using the ADASYN technique produced better results. With the imbalanced dataset (A1), it can be observed that accuracies of the base classifiers appear almost the same, though J48 is better by about one unit. The accuracy score of NB increased by more than 11%; that of SMO by about 6%; and the J48 (DT) recorded the highest increase of about 17%. It is also interesting to see how precision scores of all the classifiers significantly improved due to data balancing, with SMO showing 100% score followed by J48. Overall, it can be observed that J48, which is the chosen base classifier for our model, recorded best scores in most of the compared results. Due to ADASYN's ability to predict which spots in the minority class would be the most challenging for a model to learn, a higher ratio of synthetic data is placed adjacent to these points, which has improved performance.

Next, we show the evaluation results of our proposed Random Forests-based model (GTMP) and compare it with the other approaches. For the purpose of graphical presentation of the results, 'BaNB', 'BaSMO', and 'BaDT', were used to, respectively, denote the Bagging with Naïve Bayes, Bagging with SMO, and Bagging with Decision Tree. Also, 'BoNB', 'BoSMO', 'BoDT' and RT represent Boosting with Naïve Bayes, Boosting with SMO, Boosting with Decision Tree and Random Tree, respectively. Fig.2 to Fig.5 presents the accuracy, precision, recall and F-measure scores respectively. Figure 2 illustrates accuracy scores of the GTPM and the compared ensemble classifiers. Keep in mind that the Accuracy metric evaluates how accurately the classifier predicts both positive and negative instances. It is calculated as the proportion of true predictions to all predictions. From the figure, the highest accuracy is achieved by the GTPM with 83.74% accuracy. On the other hand, the lowest accuracy score of 59.98% is recorded by the Bagging ensemble based on the SMO algorithm. Also, it can be vividly observed from the graph that the DT base classifier resulted in better results in all the ensemble methods compared to the other base classifiers. In other words, the Bagging ensemble produced better results with DT, than both NB and SMO, with 74.9% accuracy score. Similarly, the Boosting ensemble method with DT recorded 66.87%, which is better than both NB and

SMO. Also, the Random Tree method with DT recorded 81.69% accuracy, which is better than both Bagging and Boosting methods. Lastly, the GTPM, which is also DT-based Random Forests, achieved best results compared to all others. Overall, we can conclude that, with respect to accuracy, the best base classifier is the DT and the best ensemble method is the Random Forests used in the GTPM.

The Fig. 3 shows the Precision score of the compared classifiers. The proportion of accurately anticipated positive values to all correctly predicted positive values is used to calculate precision. From Figure 2, all the classifiers have recorded good results ranging from more than 81% to about 100%. The best score goes for the SMO based Bagging ensemble with about 100% precision. This is followed by the GTPM (DT based Random Forest) classifier with 87.8% and then DT based Boosting ensemble with 87.2%. Recall that the SMO based Bagging, which records the best results here, scored the worst result in the accuracy metric. This demonstrates that the classifier could not trade-off between the two performances. The best trade-off has been achieved by the GTPM, which recorded the best results in accuracy and second best in precision metric.

The Figure 4 displays results of the different ensemble classifiers in terms of the Recall measure. The ratio of genuine positives to real positives plus negatives is used to calculate Recall, also known as Sensitivity. Real positives in this study are postgraduate students who graduated on schedule, whereas false negatives are students who were incorrectly classified as not graduating on time. Figure 3 compares the Recall ratings of the potential classifiers. From the results, the GTPM recorded the best score with 85.4% while SMO based Bagging has the lowest performance with about 60%. It can also be observed that DT appears the best base classifier for all the ensembles, noting that the GTPM is also based on DT.

The results of F-measure are depicted in Figure 5. From the results, all the classifiers have recorded good results ranging from more than 71% to about 87%. The GTPM (DT based Random Forests) still outperformed all other ensembles with a score of 86.6% and the lowest score of 71.2% was recorded by the SMO based Boosting. Here again, the DT base classifier outperformed the other base classifiers across all the ensembles.

Discussion

For students, graduating from University at the stipulated time is a significant achievement. However, there are a number of reasons why students could not complete their degrees in the anticipated amount of time. Therefore, predicting timely graduation is crucial since it aids in developing interventions to support students who are at risk of not finishing on time. An effective method for predicting student graduation is to use an ensemble of random forests. This algorithm is a great option for forecasting timely graduation since it enables the simultaneous assessment of several variables. It provides greater prediction accuracy. It can reduce data bias and provides enhanced robustness against overfitting.

In this research work we found out that Adaptive synthetic minority oversampling technique (ADASYN) is an excellent way of resolving class imbalance as it produces synthetic instances of minority class which has increased the ability of the data mining classifiers to accurately estimate when students will graduate. It was also discovered that ensemble learning performed better than the individual classifiers on decision trees. This is consistent with the work of (Finlay, 2011) which found out that ensemble learning produces more accurate results. Furthermore, the applied ensemble classifiers produced interesting performance results for the imbalance data classification problem. Nevertheless, it was found that the GTPM outperformed the other ensemble methods.

CONCLUSIONS

Students' data collected by universities should not just be for record keeping. A lot of research can be done to identify how to make good use of such available information to help students do better academically. This research collected and studied 972 students' dataset of Bayero University Kano, using classifiers ensemble technique to develop an improved students' graduation time prediction data mining model. The dataset collected for this research was not meant for data mining and thus it had to be appropriately pre-processed to make it suitable. We noticed class imbalance with the dataset as 392 instances belong to the minority category, which is our target demographic and 580 instances belong to the dominant class. The class imbalance problem was resolved using ADASYN sampling technique in order to avoid misleading performance results of the models. The Decision Tree based Random Forests ensemble was applied and a GTPM model proposed. The proposed model was evaluated and compared with other ensemble methods such as bagging, boosting and random tree. The results of this study further affirms that resolving data imbalance problems in a dataset using adaptive synthetic sampling improves performances of the prediction models. The results also showed that the proposed GTPM can predict students' graduation with considerable output quality. From the results, we conclude that, given the scenario considered in this research, PS_GTPM can be used to forecast whether a thesis-based graduate study shall be completed on-time or not. As future work there is the need to investigate other factors which were not present in the data used in this research such as employment status, departmental factors, strength of the faculty, and supervisor factors among others. This research can also be extended for other postgraduate programs.

ACKNOWLEDGEMENT

The authors wish to thank the School of Postgraduate Studies, Bayero University Kano, for providing the data for this research.

REFERENCES

Agbonlahor, O. (2022). Multilevel Analysis of Factors Predicting International Doctoral Students' Time-to-Degree Completion. Journal of Graduate Education Research, 3(1), 7. https://scholarworks.harding.edu/jger/vol3/iss1/7/

Ahmed, S., Mahbub, A., Rayhan, F., Jani, R., Shatabda, S., & Farid, D. M. (2017, December). Hybrid methods for class imbalance learning employing bagging with sampling techniques. In *2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)* (pp. 1-5). IEEE. 10.1109/CSITSS.2017.8447799

Amida, A., Algarni, S., & Stupnisky, R. (2020). Testing the relationships of motivation, time management and career aspirations on graduate students' academic success. *Journal of Applied Research in Higher Education*. https://doi.org/10.1108/JARHE-04-2020-0106

Baashar, Y., Hamed, Y., Alkawsi, G., Capretz, L. F., Alhussian, H., Alwadain, A., & Al-amri, R. (2022). Evaluation of postgraduate academic performance using artificial intelligence models. *Alexandria Engineering*

Journal, *61*(12), 9867-9878. https://doi.org/10.1016/j.aej.2022.03.021

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of educational data mining*, *1*(1), 3-17. https://doi.org/10.5281/zenodo.3554657

Breiman, L. (2004). Consistency for a simple model of random forests. Statistical Department. *University of California at Berkeley. Technical Report,(670)*. https://www.stat.berkeley.edu/~breiman/RandomForests/consistencyRFA.pdf

Brennan, J. (2019, December 10). Dealing with imbalanced Data. *Digital Catapult*. https://medium.com/digital-catapult/dealing-with-imbalanced-data-8b21e6deb6cd

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, *210*(2), 368-378. https://doi.org/10.1016/j.ejor.2010.09.029

Gareth, James; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert (2015). An Introduction to Statistical Learning. New York: Springer. pp. 315. ISBN 978-1-4614-7137-0. https://link.springer.com/book/10.1007/978-1-0716-1418-1

Gbolagade, M. D., Hambali, M. A., & Akinyemi, A. A. (2015). Predicting postgraduate performance using resample preprocess algorithm and artificial neural network. *African Journal of Computing & ICT*, *8*(1), 145-158. https://afrjcict.net/wp-content/uploads/2017/08/vol-8-no-1-issue-2-may-2015.pdf

Goenner, C. F., & Snaith, S. M. (2004). Predicting graduation rates: An analysis of student and institutional factors at doctoral universities. *Journal of College Student Retention: Research, Theory & Practice*, *5*(4), 409-420. https://doi.org/10.2190/LKJX-CL3H-1AJ5-WVPE

Hadi, N. U., & Muhammad, B. (2019). Factors Influencing Postgraduate Students' Performance: A high order top down structural equation modelling approach. *Educational Sciences: Theory & Practice*, *19*(2). https://doi.org/10.12738/estp.2019.2.004

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 1-54. https://doi.org/10.1186/s40537-019-0192-5

Kalmegh, S. (2015). Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news. *International Journal of Innovative Science, Engineering & Technology*, *2*(2), 438-446. https://ijiset.com/vol2/v2s2/IJISET_V2_I2_63.pdf

Knutson, R. (2020). Knutson, R. (2020). *Demographic and Academic Factors that Predict Degree Attainment for STEM Masters' Students at a Midwestern Public University* (Doctoral dissertation, University of South Dakota). https://www.proquest.com/openview/118132ec36bef65cedb6d15f64764a0c/1?pq-origsite=gscholar&cbl=18750&diss=y

Muthukrishnan, P., Sidhu, G. K., Hoon, T. S., Narayanan, G., & Fook, C. Y. (2022). Key Factors Influencing Graduation on Time Among Postgraduate Students: A PLS-SEM Approach. *Asian Journal of University Education (AJUE)*, *18*(1). https://doi.org/10.24191/ajue.v18i1.17169

Ngozi, A., & Kayode, O. G. (2014). Variables attributed to delay in thesis completion by postgraduate students. *Journal of Emerging Trends in Educational Research and Policy Studies*, *5*(1), 6-13. https://hdl.handle.net/10520/EJC150461

Nisbet, R., Elder, J., & Miner, G. D. (2009). *Handbook of statistical analysis and data mining applications*. Academic press. https://doi.org/10.1016/B978-0-12-374765-5.X0001-0

Olakulehin, F. K., & Ojo, O. D. (2008). Factors influencing the completion of dissertations by students of Post-Graduate Diploma in Education (PGDE) by distance learning in South-western Nigeria. *The journal for open and distance education and educational technology*, 4(1), 37-41. https://doi.org/10.12681/jode.9722

Osmanbegovic, E., & Suljic, M. (2012). Data mining approach for predicting student performance. *Economic Review: Journal of Economics and Business*, *10*(1), 3-12. http://hdl.handle.net/10419/193806

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, *6*(3), 21-45. 10.1109/MCAS.2006.1688199

Sasaki, Y. (2007). The truth of the F-measure. *Teach tutor mater*, *1*(5), 1-5. https://nicolasshu.com/assets/pdf/Sasaki_2007_The%20Truth%20of%20the%20F-measure.pdf

Shariff, S. S. R., Rodzi, N. A. M., Rahman, K. A., Zahari, S. M., & Deni, S. M. (2016, October). Predicting the "graduate on time (GOT)" of PhD students using binary logistics regression model. In *AIP Conference Proceedings* (Vol. 1782, No. 1, p. 050015). AIP Publishing LLC. https://doi.org/10.1063/1.4966105

Suhaimi, N. M., Abdul-Rahman, S., Mutalib, S., Hamid, N. H. A., & Ab Malik, A. M. (2019, August). Predictive Model of Graduate-On-Time Using Machine Learning Algorithms. In *International Conference on Soft Computing in Data Science* (pp. 130-141). Springer, Singapore. https://doi.org/10.1007/978-981-15-0399-3_11

Suhaimi, N. M., Abdul-Rahman, S., Mutalib, S., Abdul Hamid, N. H., & Malik, A. M. A. (2019). Review on Predicting Students' Graduation Time Using Machine Learning Algorithms. *International Journal of Modern Education & Computer Science*, *11*(7). 10.5815/ijmecs.2019.07.01.

Tampakas, V., Livieris, I. E., Pintelas, E., Karacapilidis, N., & Pintelas, P. (2018, June). Prediction of students' graduation time using a two-level classification algorithm. In *International Conference on Technology and Innovation in Learning, Teaching and Education* (pp. 553-565). Springer, Cham. https://doi.org/10.1007/978-3-030-20954-4_42

Thakar, P., & Mehta, A. (2015). Performance analysis and prediction in educational data mining: A research travelogue. *arXiv preprint arXiv:1509.05176*. https://doi.org/10.48550/arXiv.1509.05176

Yavuz, Ü. N. A. L., Sağlam, A., & Kayhan, O. (2019). Improving classification performance for an imbalanced educational dataset example using SMOTE. *Avrupa Bilim ve Teknoloji Dergisi*, 485-489. https://doi.org/10.31590/ejosat.638608

Verostek, M., Miller, C. W., & Zwickl, B. (2021). Analyzing admissions metrics as predictors of graduate GPA and whether graduate GPA mediates Ph. D. completion. *Physical Review Physics Education Research*, *17*(2), 020115. https://doi.org/10.1103/PhysRevPhysEducRes.17.020115