International Journal on Semantic Web and Information Systems
Volume 18 • Issue 1

Predictive Model Using a Machine Learning Approach for Enhancing the Retention Rate of Students At-Risk

Hani Brdesee, King Abdulaziz University, Jeddah, Saudi Arabia*
Wafaa Alsaggaf, King Abdulaziz University, Jeddah, Saudi Arabia
Naif Aljohani, King Abdulaziz University, Jeddah, Saudi Arabia
Saeed-Ul Hassan, Manchester Metropolitan University, Manchester, UK

ABSTRACT

Student retention is a widely recognized challenge in the educational community to assist the institutes in the formation of appropriate and effective pedagogical interventions. This study intends to predict the students at risk of low performance during an on-going course, those at risk of graduating late than the tentative timeline. and predicts the capacity of students in a campus. The data constitutes of demographics, learning, academic, and education-related attributes that are suitable to deploy various machine learning algorithms for the prediction of at-risk students. For class balancing, synthetic minority over sampling technique is also applied to eliminate the imbalance in the academic award-gap performances and late/timely graduates. Results reveal the effectiveness of the deployed techniques with long short-term memory (LSTM) outperforming other models for early prediction of at-risk students. The main contribution of this work is a machine learning approach capable of enhancing the academic decision-making related to student performance.

KEYWORDS

Academic Performance Prediction, Balancing Student Data, Classification, Early Student Prediction, LSTM, Machine Learning, Student at Risk

1. INTRODUCTION

Despite the recent growth of online education under the paradigm of Technology-enhanced learning or TEL (Cheng et al., 2021; Waheed et al., 2020; Rajabi & Greller, 2019), traditional universities remain the primary source of education for the masses. Academic performance improvement and early intervention of at-risk students' remains a challenging task in any educational setting (Fayoumi & Hajjar, 2020). Our research provides prediction-based models on the data taken from student information system by tapping the power of machine learning (Jiang, Gradus, and Rosellini 2020) to predict the academic performances of students with high accuracy and those at-risk of graduating late. The employed models assist instructors in forming appropriate pedagogical intervention strategies for optimal resource allocation of an institute (Maheshwari et al. 2020). Overall, the increased recognition of online learning platforms has yielded a progression in the data repositories about

DOI: 10.4018/IJSWIS.299859 *Corresponding Author

students' interactions and online activities, resulting in several educational research communities (Aldowah, Al-Samarraie, and Fauzy 2019; Avella et al. 2016).

In contrast to the online learning data repositories and their analysis in predicting students' academic performances, little work has been conducted on the students' interactions and influential attributes impacting their academic performances in traditional classroom settings (George & Lal, 2021). This lack of student engagement data, coupled with the reluctance of universities to share their data due to privacy concerns, becomes a hindrance in determining student academic performance and identifying at-risk students, especially in traditional classroom settings. In the existing studies, analyzing and predicting the performances of students has received considerable attention in the educational data mining community and hence in the newly emerging related fields, such as learning analytics, this particular objective has evolved in terms of early identifying the student at-risk of low performances, during an ongoing course (Chanlekha and Niramitranon 2018; Hassan et al. 2019).

Furthermore, another dimension prevalent in par with predicting academic performances is analyzing the student's time to graduate a degree. Learning analytics also emphasizes the optimal resource allocation of an institute for strategizing the administrative tasks, regulating performances and maintaining learning resources in higher education (Waheed et al. 2018). The capacity of an institute is also a significant predictor in analyzing the resource maintenance mechanism for more optimal allocation.

To assist the educational stakeholders in forming instructional pedagogical interventions, improving the academic performances of students and identifying the students at-risk of low performances and at-risk of graduating late from the institute, this study leverages machine learning techniques to analyze these perspectives of an institute. The research objectives addressed in this study intend to leverage the student data from traditional classroom settings for a more thorough analysis of student behavior, and are stated as follows:

- The first objective of this study is to deploy machine learning models to evaluate its effectiveness in the prediction of academic performances of students and predicting students at-risk of getting low grades in a traditional classroom setting.
- The second objective is to leverage deep learning techniques to predict the students at risk of graduating late and assess the average time students take to graduate using their course level information and course level activities.
- The final objective is to analyze the institute's capacity at a given particular time to identify the number of students enrolled in the institute for optimal resource allocation maintenance.

The rest of the work is organized as follows: In Section 2, we provide a comprehensive literature review and the critical contributions in the field are highlighted and discussed in alignment to our research model. Section 3 elaborates on our research methodology and empirical data management. Key findings, discussion and analysis of empirical data are provided in section 4, while conclusions and recommendations are attached in section 5.

2. RELATED WORK

Educational data, consisting of valuable, actionable information of students, significantly influence the predictions of identifying students who are at risk for low performance and those who graduate late. Learning analytics can help estimate functional learning patterns of students' academic performance and highlight students in need of interventions in their studies. This section is further sub-sectioned to present the existing studies using machine learning techniques to i) highlight the students at-risk of low academic performances and in need of intervention ii) identify students at-risk of graduating late in the institutes, which ultimately impacts the institute's resources and reputation.

2.1 Students At-risk of Low Performances

This section explores the studies conducted in the learning analytics discipline to predict a student's academic performance in terms of award-gap performances (pass/fail), grades prediction or cumulative grade point average (CGPAs). Hendrik and Andreas used daily student activity data from Virtual Learning Environments (VLE) to predict their success, i.e. whether they pass or fail a course (Heuer and Breiter 2018). The authors concluded that binary information (whether a student was active or inactive on a particular day) had the same predictive power as the exact number of clicks. They used four different supervised ML models and compared their results. Further, they also used K-means clustering to group students based on their daily activities. Zhang et al. (2017) analyzed the student engagement factors that were highly correlated with student academic performances. They observed several course logins, time of resource watch, and repeated resource watch as key factors influencing student marks/grades. These features were used in the Logistic Regression model to categorize students as excellent/not excellent.

Jiezhong et al. (2016) studied the impact of the social pressure surrounding a student; they observed that a student's tendency to get certification increases if his/her friend holds a certificate. They also found the student's inquisitive nature to impact the certification completion. Their 'learning effectiveness' model outperforms alternative methods in this field. Daud et al. (2017) showed that in addition to academic performance features, the inclusion of family expenditure and student's personal information outperforms existing methods in predicting if a student will complete the degree using C4.5 classifier. Shahiri and Husain (2015) provided an overview of the techniques used in the learning analytics discipline to find a student's final grade. They observed CGPAs and internal assessments as the datasets that are commonly used. Among the machine learning models, decision trees and neural networks were the most prevalent models. Jiang et al. (2014) used Logistic Regression to predict whether a learner will earn their online certificate. The authors used the first-week assignment result and learners' interaction within the MOOCs to estimate the probability of learners getting a certificate. Okubo et al. (2017) compared the performance of Recurrent Neural Networks (RNNs) to traditional regression techniques when predicting final student grades, using student clickstream data with the online institute platform. The use of RNN showed better predictive power

in contrast with other models.

Another essential aspect in predicting at-risk students is early identifying students during ongoing courses so that intervention strategies may be introduced to improve their academic performances for that course. Lu et al. (2018) used Principal Component Regression to estimate student performance using student behavior information such as quiz and assignment scores. Their model predicted students' performance when only one-third of the semester was completed. This early performance estimate can be used to target students in need of early intervention so that their fortunes can be potentially changed for the better. Willging and Johnson (2009) did thorough research to identify the factors that influence students to drop out of courses. Financial circumstances, age, Gender, quality of teaching and difficulty settling in with fellow students were amongst the key factors discovered. Further, this work was extended to uncover the reasons behind high dropout rates in online programs. Demographic features (such as age, Gender, occupation etc.) were combined with data collected from surveys to achieve this purpose. The survey covered various questions ranging from reasons the student chose to enroll in the program to the factors that might have led them to drop out, such as job responsibilities, lack of interaction with teachers, unsatisfactory assignments, and much more. The research concluded that traditional face-to-face programs and online courses have similar dropout predictors.

Hlosta, Zdrahal, and Zendulka (2017) identified at-risk students in the absence of a legacy data set, using the data from running presentation for training a predictive model. Learning patterns can be extracted from the behavior of students who have already submitted their assessments earlier. Aldowah et al. (2019) observed that specific learning analytics techniques are suited for certain learning problems, and the application of these techniques can help develop a student-focused strategy for improvement in the dropout rate. Moreno-Marcos et al. (2020) used self-regulated learning (SRL)

techniques to estimate at-risk students in self-paced online courses. They predicted the students who would not complete the course even when only 33% of it is completed. Chen, Johri, and Rangwala (2018) researched student at-risk across fields of Science, Technology, Engineering and Mathematics (STEM) courses. The rate students leave STEM majors is alarming, with a degree completion rate below 40%. They developed a survival analysis framework for the early identification of at-risk students. The results were promising and comparable to traditional ML approaches such as logistic regression, classification trees and boosting. The methodology worked well even with less semester information, with features like degree duration and GPA showing strong predictive power. Haiyang et al. (2018) proposed time series-based prediction method called Time Series Forest (TSF) algorithm. Their dataset consisted of students' activities and interactions with the learning environment. They found that as the daily data with time is increased to train the model, the model performance improves.

In the existing literature, several studies use various datasets to predict students' academic performances. The features used in each study vary from demographic information to student engagement data with the online learning platforms to using features from traditional classroom settings such as social pressure, family expenditure, and other extrinsic factors intrinsically influencing a student. A wide variety of features can be observed in the literature, impacting students' academic performance. In this study, we use a set of these attributes, consistent with the existing literature; more details are present in section 3, Data and Methodology.

2.2 Prediction of Students Graduation Time

A range of supervised and unsupervised techniques have been deployed in the existing literature to predict the time students take to graduate. Cahaya, Hiryanto, and Handhayani (2017) deployed the k-Medoids clustering algorithm to create clusters based on intracluster similarity. Overall, seven clusters were formed based on the data of nearly 250 graduate students and labelled according to the approximate time of students' graduation. There is a range present because the students have identical scores but different graduation times. Zulfa, Fadli, and Ramadhani (2019) applied supervised ML techniques to identify the graduating time of the students on the dataset from Jenderal Soedirman University. The study uses SVM for accurate and early warning for study programs, with an accuracy of 90.64%. The semester-achievement index becomes an important component in determining if the student will graduate on time or not.

Nurhuda and Rosita (2017) used neural networks to predict students' graduation time. They deployed various data mining techniques to extract useful patterns from the data. The data consisted of students' information such as grade points, cumulative semester credits, financial status, and job status of the student. Each feature was assigned points according to its importance. The architecture of the deployed neural network consisted of an input layer with five neurons, a hidden layer of five neurons and an output layer, with a 0.001 learning rate which produced the smallest MSE value of $4.38x10^{-06}$. Anderson, Boodhwani, and Baker (2019) used multiple machine learning methods to predict the graduation time of students. Similar to the existing researches, they deployed the prevalent machine learning techniques: decision trees, logistic regression, linear support vector machine, and stochastic gradient descent binary classifiers. They procured data from publicly funded universities with a diverse population. The included features were categorized into four levels, i.e. financial information, academic information, pre-admission information, and extra-curricular activities. They used 80% of data to perform 5-fold cross-validation on models to tune parameters, and the remaining 20% was used for testing purposes. The Stochastic Gradient Descent binary classifier performed better than others compared to other algorithms.

This section summarizes existing studies on predicting the graduation time taken by students. Such analysis can assist institutes in having an outlook for future students, managing the resources at the administrative level, and regularising the operations of an institute optimally. In par with these objectives, analyzing the capacity of an institute can also be beneficial in strategizing an institute's resources and providing optimal support to students in need of guidance. Therefore, we also intend

to highlight another vital aspect correlated with student academic performances and their graduation time in this study. Studies do not cohesively associate these objectives in the existing literature. We intend to focus on these three research objectives, inter-relating them cohesively to assist institutes in forming optimal pedagogical policies for their future.

3. DATA & METHODOLOGY

This section discusses the procured data, its processing, and statistics to analyze it in-depth. For preliminary analysis, demographic attributes were analyzed to observe their impact on students' performances. For a more comprehensive analysis, course-level information was included in conjunction with academic variables to analyze the impact on students' performances, grades, and graduation time. Moreover, influential attributes are included for a more comprehensive performance analysis after correlation analysis.

3.1 Dataset Overview

The dataset is acquired from the student information system of a Saudi university; for anonymity, the ethnicity of the dataset will not be revealed. Overall, the data consists of more than 3 million students' engagement, including transactional data of course registration the academic performance of over 230,000 students from the following years: 2006 to 2015. The dataset also includes the students who have graduated from the university and are currently pursuing their studies. Therefore, the dataset is sufficient to perform analysis on students that have graduated as well as ongoing students and their performances. It consists of demographics and academic-related features that are further categorized for a more thorough analysis. The university consists of four campuses categorized as Male A, Male B, Female A, Female B and several colleges differentiated on their majors. Due to class imbalance and bias, only specific colleges and campuses were selected for each objective, which will be explained further.

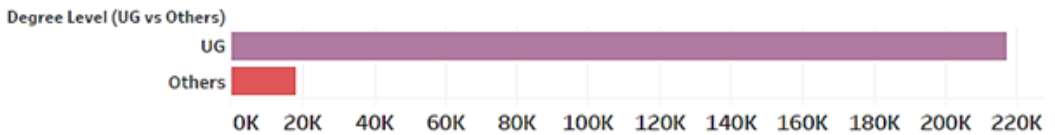
3.2 Data Summary and Statistical Analysis

The acquired dataset comprises students' demographics, academic-related information such as earned and taken credit hours, major subjects, and other degree-related information. This section presents some preliminary statistical analyses to overview the dataset.

3.2.1 Graduates Data

A distinctive class imbalance is observed in the dataset, with more than 200,000 graduate students' (Undergraduate: UG) data surpassing others (Masters, PhDs etc.), as presented in Figure 1. However, in terms of gender division, such an imbalance cannot be observed (see Figure 2), with both males and females having records above 100,000.

Figure 1. UG students versus Other Degree Level Students



The students in the dataset belonged to 4 major campuses, with two campuses for males: Male A, Male B and two campuses for females: Female A, Female B. The academic performance of female

Figure 2. Males' vs Females Count

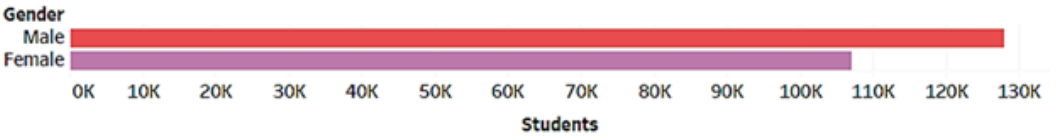


Figure 3. Median CGPA for Each Campus

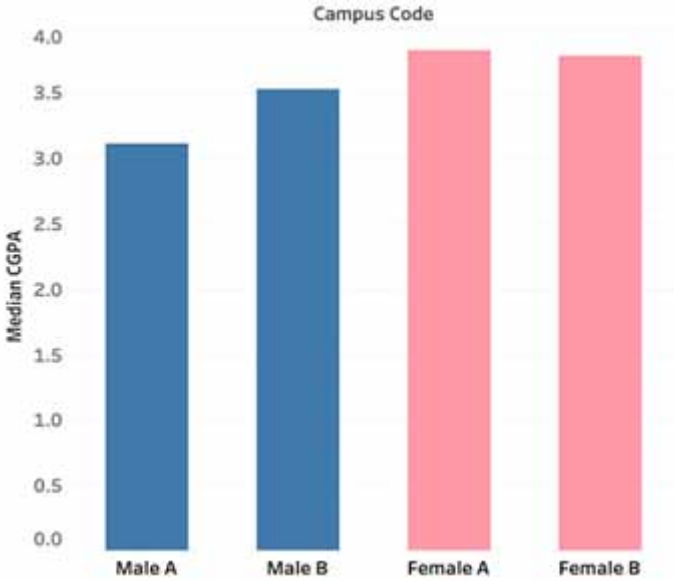
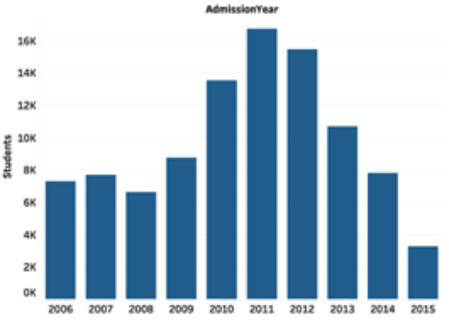
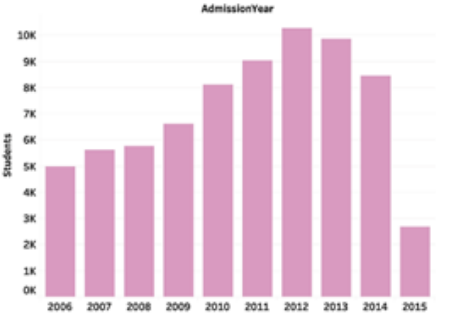


Figure 4 Graduation Rate in terms of Gender



(a) Graduation Rate of Male Students

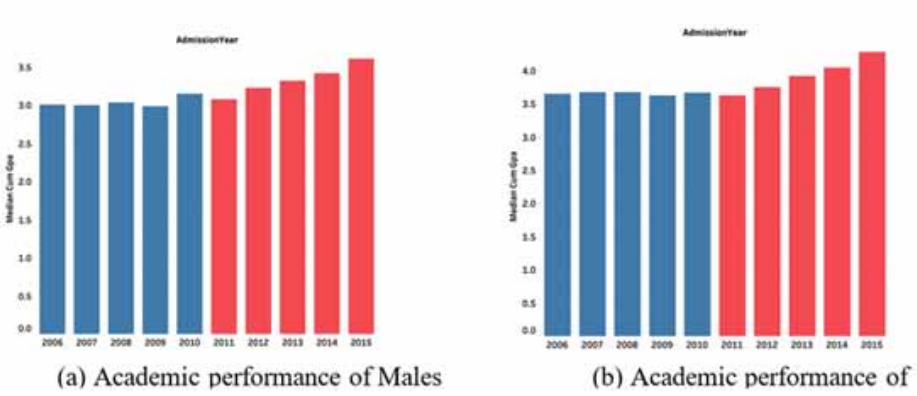


(b) Graduation Rate of Female Students

campuses was better than males, with higher median CGPAs (Figure 3), where pink bars represent female campus performances and blue bars represent male campus performances.

It was observed that the overall enrollment for both males and females tend to increase as the years' progress (Figures 4), and interestingly, their academic performance in terms of CGPA also improves (Figures 5).

Figure 5 Academic Performances of Males and Females over the Years



Most of the data for each Gender belonged to Male A and Female A campuses, respectively, with other campuses having much fewer records in comparison (Figure 6).

3.3 Data Processing

The procured data contains various features related to students' admission and learning process, including their demographical information, admission details, degree and majors selected, credit hours taken and earned information and number of years taken to complete the degree. For a more thorough analysis, the provided attributes were categorized into four sections as depicted in Figure 7: demographics, learning features, educational and academic features. Further, some preprocessing was conducted to remove any null and duplicate records.

Furthermore, only Male A campus data was used for the analysis to ensure uniformity in marking schemes and avoid any gender bias. We categorized student academic performance as either 'Good' or 'Poor', based on their CPGA. The data was provided in two sessions: one with demographical, educational, academic and learning features (without course details) and the second session with the course details. From the course details, semester information was extracted, and each student's semester subjects were extracted such that multiple records existed for one student, with each record having one subject detail and marks. Preprocessing techniques were deployed to merge the data, such as each student's information was represented in one record with their college and campus name, majors taken, credit hours taken and studied, courses studied and marked obtained in each course.

3.3.1 Identifying Students At-Risk

A two-fold analysis was conducted to identify the at-risk students of low academic performance. Firstly, classification was performed on the students' information, excluding the courses data to cater for the academic performance as good or poor. For a more thorough analysis, course level information was incorporated to measure the good and poor performances by converting it into a regression problem.

Figure 6 Distribution of students across campuses

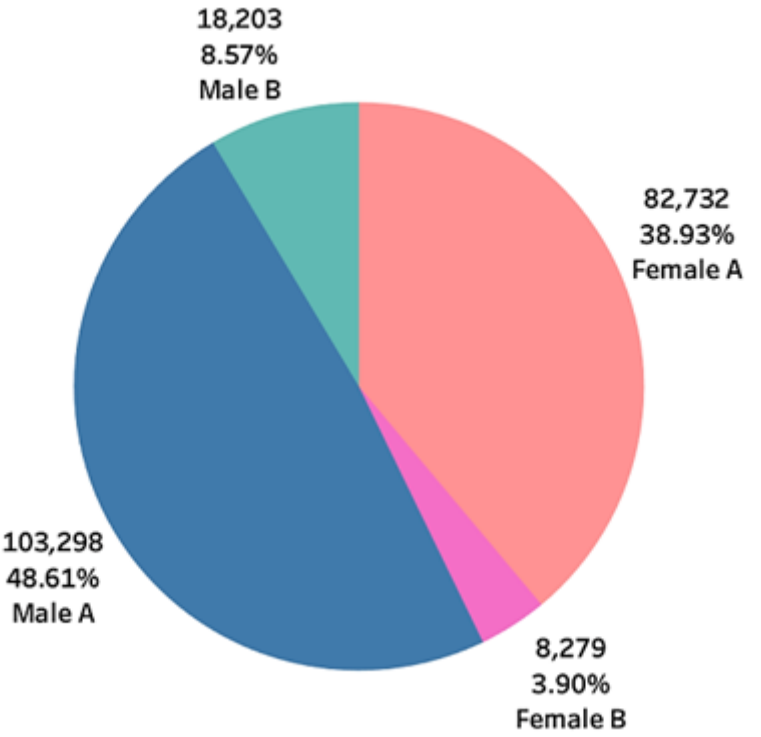


Figure 7. Data Features and Categories

Further to identify the students at risk of low performance, a combination of features were used to identify the highly correlated features impacting their performances. Pearson's correlation was applied to find the highly correlated features, and the first five highly correlated features were selected, namely: time taken to graduate, credit hours undertaken, credit hours earned, student college code and students' major taken. For a course level granularity, semester wise information was included iteratively in a sequential fashion to predict students' academic performances.

3.3.2. Graduation Time Prediction

The second objective also constituted of a two-fold analysis, where firstly a classification scheme was applied to identify the students at-risk of graduating late, and finally, course level information was incorporated to perform regression and predict the time each student takes to graduate. Semester information was mined from the available dataset from Fall/Spring term information to predict the graduation years. The provided dataset had semester information regarding years and Fall/Spring term admissions and graduations. Therefore, each student's semester information was extracted by mining their admission and graduation years, and Fall/Spring terms were used to compute the semesters for each student. In this way, the time taken to graduate was calculated based on the number of semesters.

Pearson's correlation was applied to identify the influential features strongly correlated with the graduation time. From the multiple features present, five highly correlated attributes were selected: credit hours taken, credit hours earned, major subject taken, required credit hours for that degree and the status of the student, that is, if the student was regular, had gaps in semesters or a remote learner etc. These attributes depicted a high correlation with the graduation time, where the target variable was labeled as timely and late graduates.

3.3.3. Capacity Analysis

To calculate the capacity of the campus/college at a given year, a list of the courses offered in that year and average GPA for that year were calculated, a new feature consisting of the number of students was computed from the student status and enrollment information. The capacity of the campus from the year 2012 to 2019 was computed from the dataset, and capacity for the next year was predicted from the computed information.

3.4 Modeling Approaches

This study implements a two-fold analysis for identifying the students at risk of low performance and those graduating late. Firstly processed data of the students, excluding their courses information, is used for each of the mentioned objectives and later, a semester-long analysis is performed, including the courses studied and grades obtained in each semester. This semester-long analysis will assist in early predicting students at risk of low performance and identifying those at risk of graduating late. For our first analysis, conventional machine learning algorithms were deployed, such as Logistic Regression, Decision Trees and Random Forest. The sequential Long Short Term Memory (LSTM) model was implemented for the early prediction of at-risk students. This section contains a description of each of these approaches.

3.4.1 Logistic Regression (LR)

LR is one of the most prevalent baseline methods deployed in this discipline, where it makes use of several independent variables to find the probability of a categorical dependent variable (Eckles and Stradley 2012). The equation for this classifier is defined mathematically in Eq.1.

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x \tag{1}$$

where b_0 is the constant that provides the shift towards left or right, b_1 depicts the slope outlining the gradient, and p is the logistic model, as depicted in Eq. 2:

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x)}} \tag{2}$$

LR was deployed on predicting the academic performances of students and their graduation time where the academic performances of students were categorized as good/low, and their graduation time was converted into a classification problem with the class label asserted as late/not late.

3.4.2 Decision Trees

Decision trees are flowchart type tree-like structures with leaves and nodes, where the nodes represent an attribute and leaves represent the class label. It conceptually works by splitting the data such that all the attributes are partitioned to output a leaf belonging to one class (such as low/high performance, timely/late graduate). Splitting of the nodes is based on a scoring function that determines the node purity (Kabra and Bichkar 2011). The attribute that outputs the purest node is selected. In this study, experiments were conducted using the Gini index and entropy information gain to calculate the purity of each node. Decision trees were implemented with entropy and Gini index, with the depth ranging from 3 to 6 and minimum leaves 4 to 7. With increasing depth issues of over-fitting surface, therefore, a range has to be set. From these experiments, entropy with a depth of 5 and minimum leaves of 6 was found to be the best fit amongst other decision tree experiments.

3.4.3 Random Forest

These are ensemble methods for both classification and regression, operating on a multitude of decision trees during training and deciding the resultant class on the mean prediction of individual trees. Since the results depend on multiple trees, therefore they are less prone to overfitting issues and comparatively perform well than decision trees.

3.4.4 Long Short Term Memory (LSTM)

LSTMs are used for sequential time-series data where it is important to retain the previous information of the instances. It is comprised of an additional memory unit, which enables it to capture the information to be used in the future (Wang et al., 2016). In our study, LSTMs were implemented for the early prediction of at-risk students during an ongoing course. Due to their memory unit, issues of vanishing and exploding gradients are minimized in LSTMs; therefore, these have been given considerable attention (Okubo et al., 2017).

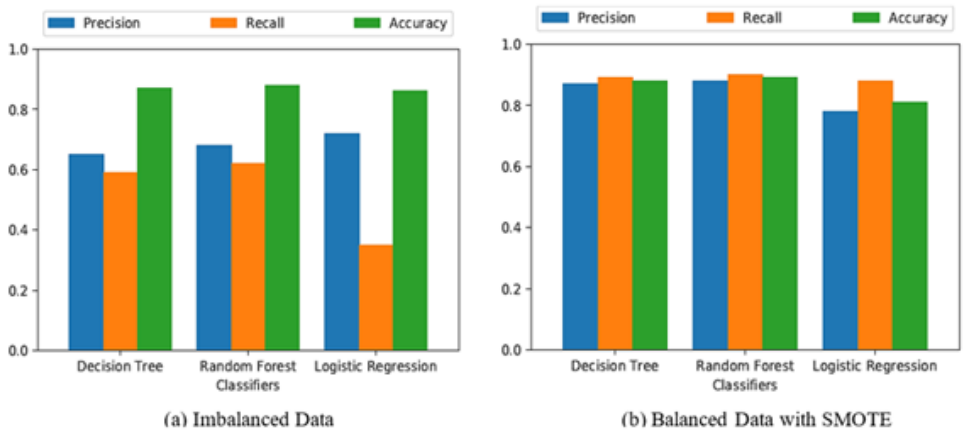
3.5 Handling Class Imbalance

For the first two objectives of our study, pertaining to the prediction of students' performances and their graduation time, an imbalance was observed in students' academic performances in terms of good/low performances and timely/late graduates. To eliminate this imbalance, Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al. 2002), prevalent in the learning analytics community, was deployed on the processed dataset. Using the concept of K nearest neighbors, it synthetically creates new minority class instances between pre-existing instances. Detailed explanations of the results derived using SMOTE have been discussed in the Results section.

3.5.1 Evaluation Metrics

Some conventional machine learning metrics were applied for the evaluation of the deployed classifiers and regressors. Since accuracy alone is not a significant measure for evaluating the performance of

Figure 8 Classifying Students At-risk of Low Performance



a model, especially in an imbalanced dataset, therefore precision, recall and area under the curve (AUC) were also used for assessing the performance of the deployed models. Accuracy denotes the percentage of the correctly classified instances by the model, and therefore in imbalanced datasets, it does not represent the actual performance of the model since it is unable to distinguish the correctly classified instances belonging to different classes. Precision and recall are defined as ratios, with the former one quantified as the number of students at-risk that are actually at-risk in the dataset and the later metric is specified as the number of the at-risk students specified by the model out of all the at-risk students in the dataset. AUC is a scale-invariant metric representing the degree of separability of the classified predictions within a range of 0 to 1. A 0.5 AUC value represents a random classifier; a value closer to 1 indicates a good classifier distinguishing between at-risk students and those who are not at risk of low performance or graduating late (Khajah, Lindsey, and Mozer 2016).

3.5.2 Training and Validation

A 10-fold cross-validation technique was used to ensure that train-test data splits remain unbiased for model evaluations (Wong 2015). This technique was applied to identify the students at risk of low performance and predict their graduation time. Such a technique will enable the model to predict without bias.

4. RESULTS & DISCUSSION

This section reports the result for each of the described objectives of this study. Chronologically, this section describes the results for identifying the students at risk of low academic performances, predicting the graduation time taken by the students and identifying late graduates, and lastly, determining the capacity of a college at a given particular time. A detailed description of the experimental setups is also discussed in this section.

4.1 Identifying Students At-Risk

To predict students' academic performances in terms of good/low performances, a two-fold analysis was conducted in the form of classification and regression, as described in section 3.2.1. For the classification analysis, student information mentioned in Figure 7 was applied, excluding the course information and details. Further for each analysis, experimental results are presented with and without SMOTE to determine the significance of this technique and evaluate the effect of class imbalance in

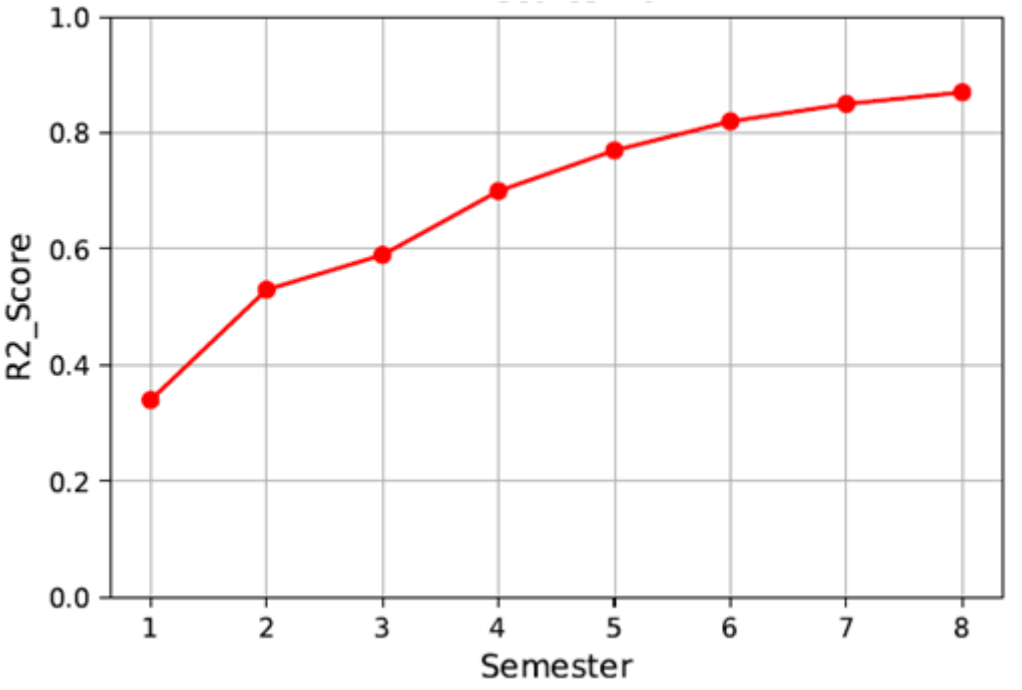
students' academic performances. Highly correlated features were included in the machine learning models, identified through Pearson's Correlation analysis. Three conventional machine learning models such as Decision Tree, Random Forest and Logistic Regression, prevalent in the learning analytics community, were deployed. Figure 8 illustrates the differences in the model predictions for students at-risk of low performance, with and without SMOTE technique. It can be observed that the results with SMOTE are better than the ones with the class imbalance. This highlights the significance of employing class balancing techniques, which enables the model to predict more robustly.

For a more thorough analysis, courses information was also included, and students' academic performance was assessed by transforming the prediction problem into a regression task. This analysis also enabled early prediction of students' performances in terms of identifying students at-risk of a low performance during an ongoing course. For each student, their courses taken were segregated semester-wise such that based on previous semester performances, next semester performances were predicted in terms of CGPA achieved by the students. The semester cutoff was gradually increased till there was a good balance between model accuracy and early detection. The students with a predicted CGPA of less than 3.0 were the one's identified in need of early intervention.

4.1.1 Early Identification of At-risk Students Using Regression

Ridge regression (Marquardt and Snee 1975) inherently uses linear regression with L2 normalization, and R2 score (Kramer 2005) was the performance metric used for model evaluation. Ridge regression was applied for initial experimentation to identify the students at risk of obtaining a low CGPA. Since highly correlated features appended with courses information were used as a feature set; therefore, ridge regression was selected as a model for prediction. Ridge regression performs better with a feature set with highly correlated features because it reduces variance among the features (Marquardt and Snee 1975). The R2 is a statistical measure used for regression tasks that gives a measure of the performance of the predictions by providing a value between 0 and 1; the higher the value of the R2

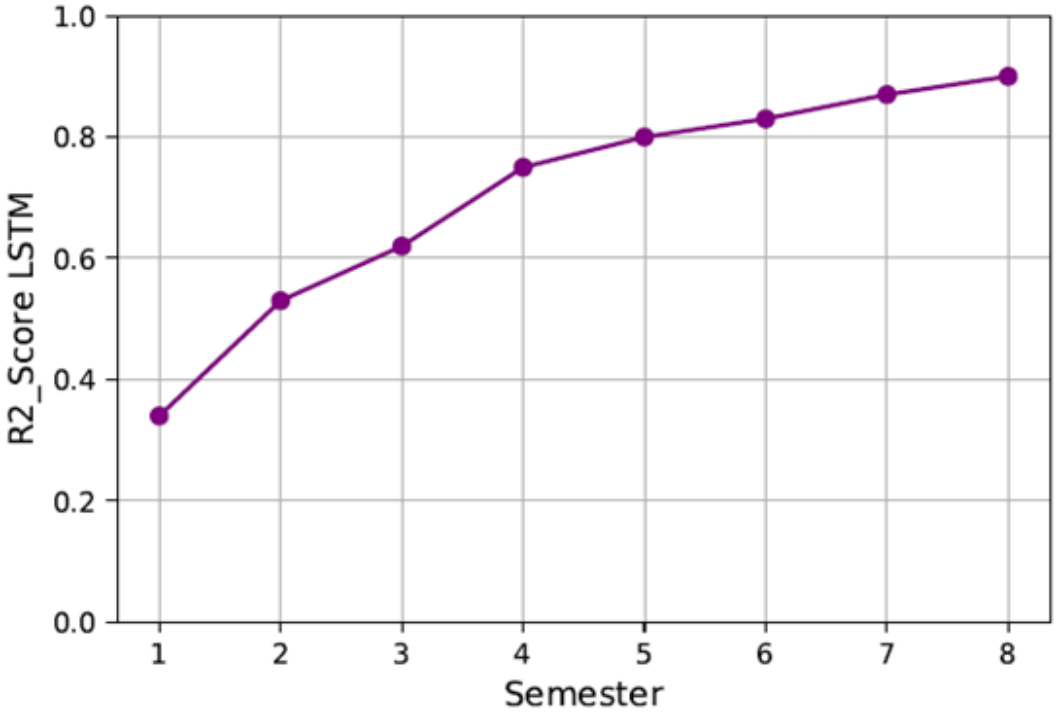
Figure 9 Identification of At-risk Students using Ridge Regression



score, the more robust the model is. It represents the best fit of the regression model, with a value of 1 indicating that the predicted regression values perfectly fit the actual data. Default values of ridge regression were implemented in python.

Figure 9 depicts the regression results achieved by the model corresponding to the cutoff semester course level information provided for training. It can be observed that as the semester information increases, the regression results are improved. For this analysis, semester wise information for each student was appended sequentially and thus; it enabled the early prediction of students at-risk of a low CGPA.

Figure 10 Identification of At-risk Students using LSTM



4.1.1.2 Improving Regression Results using LSTM

Since the data is transformed into a semester-wise sequential fashion, therefore to improve the predictions further, LSTM was deployed. LSTM works well for longer sequences, and its look-back window is flexible enough to be applied to such sequential data. Results for LSTM regression are presented in Figure 10. LSTM was applied with a look-back window for each semester depending on the number of previous semesters. Based on the previous performance history of each student, CGPA for the next semester were predicted. It can be observed in Figure 10 that as the number of semesters increases, the model gets more data to analyze the student academic performance and thus makes robust predictions. Comparing the results of Ridge regression with LSTM, it can be observed that LSTM performs better than Ridge regression, pertaining to its capability of learning from the previous history. With each semester cutoff, the predictions are improved in LSTM.

For LSTM, extensive experiments were conducted to find the optimal results for predicting the student at-risk of low performances. The deep architecture consisted of two LSTM layers with 50 units; activation 'ReLu' was observed to provide good results. A bias regularizer of 0.01, with

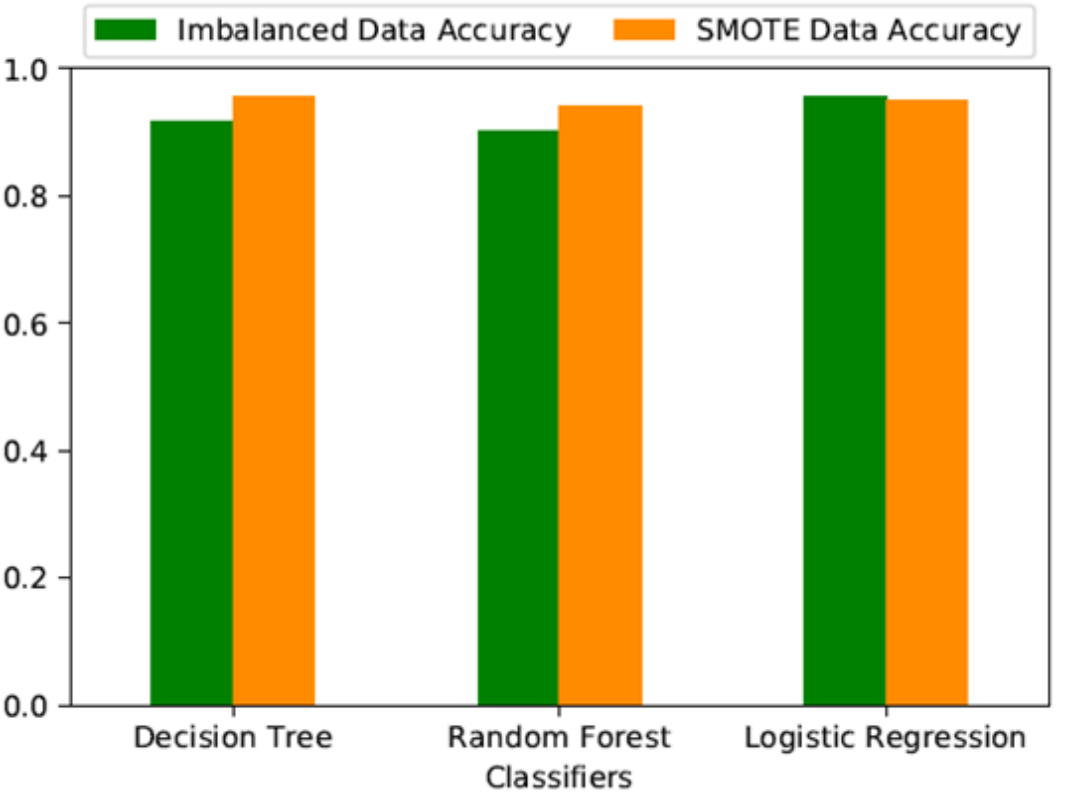
a dropout and batch normalization layer, was introduced in each LSTM layer to reduce overfitting in the model. Since the problem was a regression one, therefore mean square error was compiled in loss function, and 'Adam' was introduced as an optimizer, with a batch size of 64 and 200 epochs. Comparing Figures 9 and 10, it can be observed that LSTM predicts the academic performances with a higher R2 score as compared to ridge regression, that does not cater to the sequential information. In the 4th semester, ridge regression has a 0.65 R2 score (refer to Figure 9), whereas LSTM has a 0.68 R2 score (refer to Figure 10). Pertaining to our problem of early identifying the at-risk students, LSTM performs well comparatively with an increase of 4.61% R2 score in the 4th semester.

4.2 Graduation Time Prediction

To identify the time, a student takes to graduate, analysis was conducted where the prediction was first done as a classification task, including the learning and academic features with the inclusion of courses information from which semester-wise data was computed. Since class imbalance was observed in this task, where the number of students who graduate in 5 years is twice more than those who graduate in four years, six or more years, SMOTE was applied to balance the data.

To predict the graduation time of students in terms of binary classification as timely/late graduates, academic and learning features along with courses information were included in the feature set. Semester information was computed from the provided courses data; admission semester (spring/fall) and graduation semester (spring/fall) were used to compute the graduation time of each student. Graduation year was provided in the dataset, and the computed semester information was used to predict the graduation time of each student. Some conventional machine learning classifiers,

Figure 11. Comparison Results for Graduation Prediction in Imbalanced vs Balanced Data

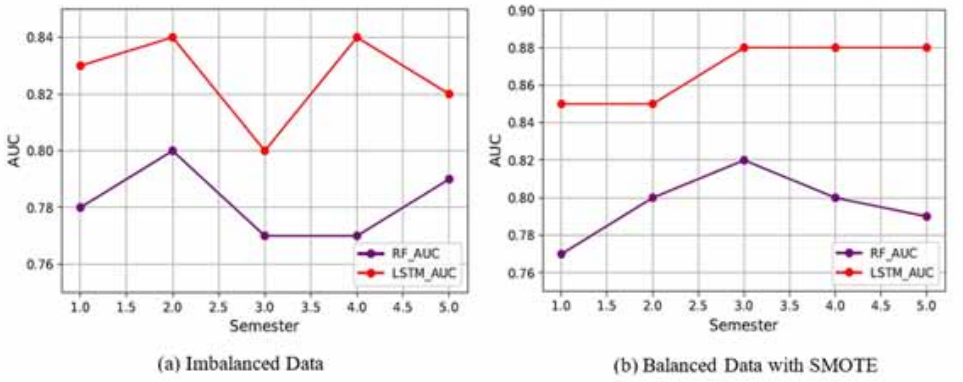


prevalent in the learning analytics community, were deployed in order to compare their performance. To eliminate the imbalance in the data, SMOTE was employed and resulted from imbalanced and balanced data were compared to observe the effectiveness of the deployed balancing technique, as depicted in Figure 11.

4.2.1. Early Prediction of Students At-risk of Graduating Late

To early predict the students at-risk of graduating late during an ongoing course, semester information was computed from the provided courses data, admission semester (spring/fall) and graduation semester (spring/fall) were used to compute the graduation time of each student. Since, in the dataset, the least number of semesters taken to complete graduation is six years; therefore information of up to 6 semesters is taken for each student. This analysis will assist in identifying the students that are at risk of graduating late during their ongoing semesters. Students at risk of graduating late were predicted using LSTM and RF for each semester cutoff for this analysis. Since LSTM is known for sequential data, each semester was iterated in the model to learn students' behaviour from previous semesters. The results from LSTM were compared with RF. Since RF does not cater for the sequential nature of the data, therefore in this model, each semester was appended with its previous semester, and each student's information about courses and semesters was computed as an open record.

Figure 12 Predicting Graduation Time of Students using LSTM and RF



For LSTM, extensive experiments were conducted to find the optimal results for predicting the graduation time of students. The deep architecture consisted of two LSTM layers with 50 units; activation 'ReLu' was observed to provide good results. A bias regularizer of 0.01, with a dropout and batch normalization layer, was introduced in each layer to reduce overfitting in the model. Since the problem was a regression one, therefore mean square error was compiled in the loss function, and 'Adam' was introduced as an optimizer, with a batch size of 64 and 200 epochs. For RF, default values were implemented in python. Figure 12a demonstrates the results are demonstrated in Figure 12a, where a comparison is presented between the AUC (area under the curve) of LSTM and RF. Since AUC is a better metric for prediction in imbalanced datasets, therefore accuracy was not considered for this analysis. Further, to eliminate the imbalance in the data existing for various graduation years, where the students graduating in 4 years superseded those passing the degree in 3 or 5 years, SMOTE was employed on the semester-wise appended data. Since there is a class imbalance in the data, therefore Figure 12a cannot be specified as the right depiction of the behavior trends existing in the data. Figure 12 b presents the result analysis of the balanced data through SMOTE on LSTM

and RF. The architecture for LSTM was kept consistent as employed on the imbalanced dataset. It can be observed in Figure 12 b that after eliminating class imbalance, results are improved for LSTM. It can also be observed from Figure 12 that 2nd semester seems to be the most important predictor for the graduation time of a student.

Further, drilling down in the courses offered in the earlier semesters reveals the crucial significance of these courses in degree completion. Therefore, due to the significant correlation of the courses offered in the earlier semesters, we infer that students' academic performances in earlier semesters enable the model to learn student behavior and have an influential impact on predicting the graduation time of students. LSTM model outperformed RF in early identification of students at-risk of graduating late. In the presence of class imbalance, LSTM produced an accuracy score of around 85% with an AUC score of around 80% in the first three semesters, while RF showed an accuracy score of around 86% but with a low AUC score. A Low AUC score of RF insinuates that the model had a hard time predicting samples from minority classes. Furthermore, we also analyzed the score improvement of accuracy and AUC after eliminating the class imbalance, using SMOTE, enabling the model for a more robust prediction of the graduation time.

4.3 Capacity Prediction

To predict the capacity of students enrolled in the campus institutes for a particular year, the number of enrolled students were computed using the admission/enrollment year, their graduation year, and status information, whether a student is active or regularly taking classes or has been dropped off. The number of students residing for a particular year was computed from these features. This enabled us to form a list of the count of students for each year. Capacity for a year can thus be formulated through the following formula:

Capacity = Admission (Student) £ Current (Year) &
Graduation (Student) ^{3} Current (Year) &
Status (Student) = Regular | Active | Dropout

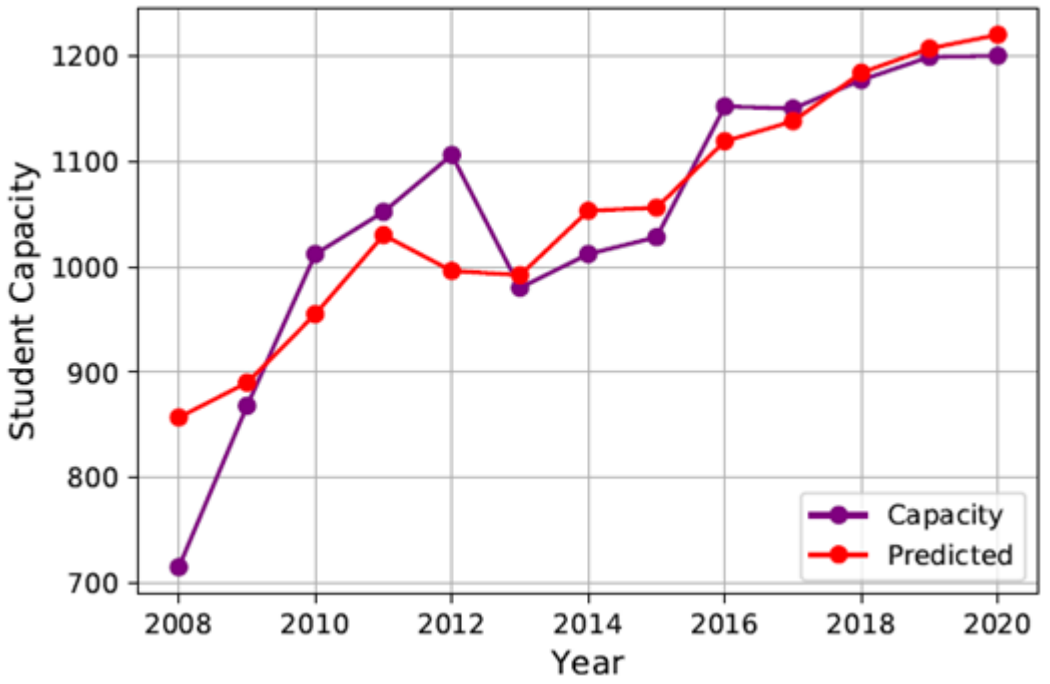
Using the additional data of students' enrollment of the selected university, student capacity analysis was computed for the provided years from 2008 to 2020. Student capacity data points were predicted for the each year using the linear regression model. As illustrated in Figure 13, a comparison is presented of the capacity computed from data and the ones predicted from the linear regression model. It can be observed that the model has performed well with an R2 score of 0.718. An R2 score value above 0.5 depicts the model to distinguish between actual and predicted values and presents the goodness fit of the model.

5. IMPLICATIONS & CONCLUSIONS

This study presented an analysis on the student's data from a government university, transforming that data into an actionable format to infer student behavior; predict their academic performances in terms of low and high performances, predict the time students will take to graduate and analyze the capacity of the campuses residing in an institute. Furthermore, to identify students at risk of low performances and at risk of graduating late, semester information for each student is computed, and the trade-off between degree timestamp and evaluation metrics is also presented. Diverse experimentation is conducted for each analysis, ranging from traditional ML models such as Decision Trees to the more advanced deep learning techniques, namely LSTMs. The study also demonstrates that a student's demographic and learning features can be effectively used to predict their academic performances. In this study, we have to use additional features to analyze the impact of these features on their academic performances. The inclusion of course-level information for each semester can be further leveraged to identify at-risk students in need of early academic intervention.

A student's academic performance prediction at the end of a course is treated as a classification problem, categorized as either low or high. Consistent with the existing literature, the performance of three machine learning algorithms (Decision Tree, Random Forest, and Logistic Regression) is

Figure 13 Actual vs Predicted Student Capacity in a Campus



compared on balanced and imbalanced data sets. RF classifier seems to perform the best amongst others, slightly better than the decision tree. The application of SMOTE oversampling technique, deployed to eliminate the class imbalance, improves the overall capabilities of all the models. Early identification of at-risk students is treated as a regression problem, with predicted CGPA underlining the severity of academic intervention required. We start with Linear Regression and feed course level information to the model up to a fixed number of semesters. Consistent with the existing literature, model accuracy improves as more and more semester data is incorporated for the training set. To further improve our results, the dataset is taken as a time sequence, with each sequence representing a semester's information, and LSTM is deployed for the early prediction of at-risk students.

For our second analysis, learning and academic features (credit hours earned, credit hours taken, campus code, major code, etc.) are used to predict a student's timely graduation. Performances of three machine learning models (Decision Tree, Random Forest and Logistic Regression) are compared on a balanced and imbalanced dataset with respect to AUC score. After balancing data with SMOTE, AUC scores are increased for all models. For the decision tree, AUC increases from 0.91 to 0.96; for RF and logistic regression, there is a slight increase in the AUC score for the balanced class. Further, course-level information is also incorporated to compute semester-wise student data for early prediction of timely and late graduates. The course level information includes courses information and GPA obtained in each course. LSTM is employed for this analysis, and with each semester cutoff, the model improves its performance.

Lastly, the study presents capacity analysis to analyze the number of students present on the campus in a given year and their prediction for the future years. Information about students' admission, graduation year, and status is taken to compute the capacity for a particular year. A student enrolled in the previous years, pursuing their graduation degree is also counted in the capacity number for each subsequent year. Linear regression is deployed to predict capacity for each subsequent year with an R2 score of 0.718. Such studies could potentially increase the universities graduation rate

by targeting the students who are best in need of intervention. It can also help higher education to make a strong decision for sustainable education. Instructors can promote loyalty and trust in their subjects by focusing on identified students and catering to their needs. Moreover, it can also assist institutes to prepare themselves for the following years by regulating their operations and resources and maintaining their principles, specifically for the years where they have abundant admissions. Capacity analysis can assist educationalists to enhance an institute performance and optimally strategizing its resources and operations for an upcoming semester.

In the existing literature, there is a lack of such descriptive studies that so deeply explore this dimension of student behavior, including the class imbalance in academic performance prediction and exploring the capacity analysis of an institute to form optimal policies for future prospects. Future avenues should explore more dimensions of student behavior impacting their academic performances, especially in class balancing issues. For sequential classification problem, class balancing is still a novel problem that needs to be addressed by the learning analytics community. Sequential machine learning models such as LSTM are still incapable of handling the upsampling of students in a temporal setting. Therefore, techniques to handle the upsampling of temporal educational data should be explored, assisting institutes in the formation of pedagogical interventions, building early alarm systems for student retention, and maintaining mechanisms for proper resource allocations.

Such studies can assist administrative authorities and educational stakeholders in streamlining formative pedagogical guidelines for instructors and institutions, identifying the at-risk students of failure, and intervening in a timely manner to offer support and guidance. The impact of such studies can be accentuated with the formation of appropriate support groups and instructional committee cells that work for the betterment of students by providing them with suitable pedagogical interventions and ultimately devising corrective strategies to enhance students' academic performances.

Supplementary Materials: Not applicable

Author Contributions: All authors contributed equally to all the phases of the research study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to institutional policies

ACKNOWLEDGMENT

Authors would like to thank the Deanship of Admission and Registration King Abdulaziz University, Jeddah for the support.

FUNDING

This research was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, Saudi Arabia, Funding No: G: 584-156-1438.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational Data Mining and Learning Analytics for 21st Century Higher Education: A Review and Synthesis. *Telematics and Informatics*, *37*, 13–49. doi:10.1016/j.tele.2019.01.007

Anderson, H., Boodhwani, A., & Baker, R. (2019). Predicting Graduation at a Public R1 University. *Proceedings of the 9th International Learning Analytics and Knowledge Conference*.

Avella, J. T., Kebritchi, M., Nunn, S. G., & Kanai, T. (2016). Learning Analytics Methods, Benefits, and Challenges in Higher Education: A Systematic Literature Review. *Online Learning*, *20*(2), 13–29.

Cahaya, L., Hiryanto, L., & Handhayani, T. (2017). Student Graduation Time Prediction Using Intelligent K-Medoids Algorithm. In *2017 3rd International Conference on Science in Information Technology (ICSITech)*. IEEE. doi:10.1109/ICSITech.2017.8257122

Chanlekha, H., & Niramitranon, J. (2018). Student Performance Prediction Model for Early-Identification of at-Risk Students in Traditional Classroom Settings. *Proceedings of the 10th International Conference on Management of Digital EcoSystems*. doi:10.1145/3281375.3281403

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Philip Kegelmeyer, W. (2002). SMOTE: Synthetic Minority over-Sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. doi:10.1613/jair.953

Chen, Y., Johri, A., & Rangwala, H. (2018). Running out of Stem: A Comparative Study across Stem Majors of College Students at-Risk of Dropping out Early. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. ACM. doi:10.1145/3170358.3170410

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017). Predicting Student Performance Using Advanced Learning Analytics. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee. doi:10.1145/3041021.3054164

Eckles, J. E., & Stradley, E. G. (2012). A Social Network Analysis of Student Retention Using Archival Data. *Social Psychology of Education*, *15*(2), 165–180. doi:10.1007/s11218-011-9173-z

Haiyang, L., Wang, Z., Benachour, P., & Tubman, P. (2018). A Time Series Classification Method for Behaviour-Based Dropout Prediction. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*. IEEE. doi:10.1109/ICALT.2018.00052

Hassan, S.-U., Waheed, H., Aljohani, N. R., Ali, M., Ventura, S., & Herrera, F. (2019). Virtual Learning Environment to Predict Withdrawal by Leveraging Deep Learning. *International Journal of Intelligent Systems*, *34*(8), 1935–1952. doi:10.1002/int.22129

Heuer, H., & Breiter, A. (2018). Student Success Prediction and the Trade-Off between Big Data and Data Minimization. In D. Krömker & U. Schroeder (Eds.), *DeLFI 2018 - Die 16. E-Learning Fachtagung Informatik. Gesellschaft für Informatik e.V.*

Hlosta, M., Zdrahal, Z., & Zendulka, J. (2017). Ouroboros: Early Identification of at-Risk Students without Models Based on Legacy Data. Proceedings of Seventh International Learning Analytics & Knowledge Conference. doi:10.1145/3027385.3027449

Jiang, S., Williams, A., Schenke, K., Warschauer, M., & O'dowd, D. (2014). Predicting MOOC Performance with Week 1 Behavior. *The 7th International Conference on Educational Data Mining*.

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised Machine Learning: A Brief Primer. *Behavior Therapy*, *51*(5), 675–687. doi:10.1016/j.beth.2020.05.002 PMID:32800297

Kabra, R. R., & Bichkar, R. S. (2011). Performance Prediction of Engineering Students Using Decision Trees. *International Journal of Computers and Applications*, *36*(11), 8–12.

Kramer, M. (2005). R2 Statistics for Mixed Models. *Proceedings of the Conference on Applied Statistics in Agriculture*, *17*.

Lu, Huang, Huang, Lin, Ogata, & Yang. (2018). Applying Learning Analytics for the Early Prediction of Students' Academic Performance in Blended Learning. *Journal of Educational Technology & Society, 21*(2), 220–32.

Maheshwari, Roy, Pandey, & Rautray. (2020). Prediction of Factors Associated with the Dropout Rates of Primary to High School Students in India Using Data Mining Tools. In *Frontiers in Intelligent Computing: Theory and Applications*. Springer.

Marquardt, D. W., & Snee, R. D. (1975). Ridge Regression in Practice. *The American Statistician*, *29*(1), 3–20.

Moreno-Marcos, P. M., Muñoz-Merino, P. J., Maldonado-Mahauad, J., Mar Pérez-Sanagustín, C. A.-H., & Kloos, C. D. (2020). Temporal Analysis for Dropout Prediction Using Self-Regulated Learning Strategies in Self-Paced MOOCs. *Computers & Education*, *145*, 103728. doi:10.1016/j.compedu.2019.103728

Nurhuda, A., & Rosita, D. (2017). Prediction Student Graduation on Time Using Artificial Neural Network on Data Mining Students STMIK Widya Cipta Dharma Samarinda. *Proceedings of the 2017 International Conference on E-commerce, E-Business and E-Government*. doi:10.1145/3108421.3108431

Okubo, F., Yamashita, T., Shimada, A., & Shin'ichi, K. (2017). Students' Performance Prediction Using Data of Multiple Courses by Recurrent Neural Network. *Proc. ICCE2017* 439–44.

Qiu, J., Tang, J., Liu, T. X., Gong, J., Zhang, C., Zhang, Q., & Xue, Y. (2016). Modeling and Predicting Learning Behavior in MOOCs. In *Proceedings of the ninth ACM international conference on web search and data mining*. ACM. doi:10.1145/2835776.2835842

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414–422. doi:10.1016/j.procs.2015.12.157

Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, *104*, 106189. doi:10.1016/j.chb.2019.106189

Waheed, H., Hassan, S.-U., Aljohani, N. R., & Wasif, M. (2018). A Bibliometric Perspective of Learning Analytics Research Landscape. *Behaviour & Information Technology*, *37*(10–11), 941–957. doi:10.1080/0144929X.2018.1467967

Wang, P., Song, Q., Han, H., & Cheng, J. (2016). Sequentially Supervised Long Short-Term Memory for Gesture Recognition. *Cognitive Computation*, *8*(5), 982–991. doi:10.1007/s12559-016-9388-6

Willging, P. A., & Johnson, S. D. (2009). Factors That Influence Students' Decision to Dropout of Online Courses. *Journal of Asynchronous Learning Networks*, *13*(3), 115–127.

Wong, T.-T. (2015). Performance Evaluation of Classification Algorithms by K-Fold and Leave-One-out Cross Validation. *Pattern Recognition*, *48*(9), 2839–2846. doi:10.1016/j.patcog.2015.03.009

Zhang, W., Huang, X., Wang, S., Shu, J., Liu, H., & Chen, H. (2017). Student Performance Prediction via Online Learning Behavior Analytics. In *2017 International Symposium on Educational Technology (ISET)*. IEEE. doi:10.1109/ISET.2017.43

Zulfa, M. I., Fadli, A., & Ramadhani, Y. (2019). Classification Model for Graduation on Time Study Using Data Mining Techniques with SVM Algorithm. In AIP Conference Proceedings (Vol. 2094). AIP Publishing LLC. doi:10.1063/1.5097475

Hani Brdesee is an Associate Professor in Information Systems (IS), Electronic Business and E-trends, and associated with the Computer and Information Technology Department, Faculty of Applied Studies, King Abdulaziz University (KAU), Jeddah, Saudi Arabia. He received his Ph.D. degree in Information Systems from RMIT University, Australia. He is a Vice-dean of the Deanship of Admission and registration, KAU, and as a General Supervisor of the University Academic Departments Heads Forum (HSAD).

Wafaa Alsaggaf is an Assistant Professor at the Faculty of Computing and Information Technology in King Abdul Aziz University, Jeddah, Saudi Arabia. She holds a PhD in Computer Science from RMIT University, Australia. Her research interests are in the areas of mobile and ubiquitous learning, educational technologies, computer science education, and machine learning.

Naif Aljohani is an Associate Professor at the Faculty of Computing and Information Technology in King Abdul Aziz University, Jeddah, Saudi Arabia. He holds a PhD in Computer Science from the University of Southampton, UK. He received the Bachelor's degree in Computer Education from King Abdul Aziz University, 2005. In 2009, he received the Master degree in Computer Networks from La Trobe University, Australia. His research interests are in the areas of mobile and ubiquitous computing, mobile and ubiquitous learning, learning and knowledge analytic, semantic web, Web Science, technology enhanced learning and human computer interaction.

Saeed Ul Hassan is an Associate Professor in AI/Data Science in the Department of Computing and Mathematics at Manchester Metropolitan University – United Kingdom - a former Senior Research Fellow at the United Nations University – with more than 15 years of hands-on experience of applications of artificial intelligence to solve real-world problems and software development client work. Dr Saeed earned his PhD in Information Management from the Asian Institute of Technology. He has also served as a Research Associate at the National Institute of Informatics in Japan. Dr Saeed's research interests lie within Applied AI, Machine Learning, Scientometrics, Altmetrics and Educational Data Science. Furthermore, he has published over 75 papers in reputed international journals and conference proceedings. Dr Saeed is also the recipient of the James A. Linen III Memorial Award for his outstanding academic performance. More recently, he has been awarded Eugene Garfield Honorable Mention Award for Innovation in Citation Analysis by Clarivate Analytics, Thomson Reuters.