Nurse Education Today 99 (2021) 104784

Contents lists available at ScienceDirect

Nurse Education Today

journal homepage: www.elsevier.com/locate/nedt

Predicting nursing baccalaureate program graduates using machine learning models: A quantitative research study☆

Li Hannaford [a,*], Xiaoyue Cheng [b], Mary Kunes-Connell [a]

[a] College of Nursing, Creighton University, 2500 California Plaza, Omaha 68178, NE, USA
[b] Department of Mathematics, University of Nebraska at Omaha, 6001 Dodge Street, Omaha 68182, NE, USA

ARTICLE INFO

Keywords:
Graduation rate
Dropout risk
Machine learning
Nursing education

ABSTRACT

Background: Despite powerful efforts to maximize nursing school enrollment, schools and colleges of nursing are faced with high rates of attrition and low rates of completion. Early identification of at-risk students and the factors associated with graduation outcomes are the main foci for the studies that have addressed attrition and completion rates in nursing programs. Machine learning has been shown to perform better in prediction tasks than traditional statistical methods.

Objectives: The purpose of this study was to identify adequate models that predict, early in a students career, if an undergraduate nursing student will graduate within six college years. In addition, factors related to successful graduation were to be identified using several of the algorithms.

Design: Predictions were made at five time points: the beginning of the first, second, third, fourth years, and the end of the sixth year. Fourteen scenarios were built for each machine learning algorithm through the combinations of different variable sections and time points.

Settings: College of Nursing in a private university in an urban Midwest city, USA.

Participants: Seven hundred and seventy-three full time, first time, and degree-seeking students who enrolled from 2004 through 2012 in a traditional 4-year baccalaureate nursing program.

Methods: Eight popular machine learning algorithms were chosen for model construction and comparison. In addition, a stacked ensemble method was introduced in the study to boost the accuracy and reduce the variance of prediction.

Results: Using one year of college academic performance, the graduation outcome can be correctly predicted for over 80% of the students. The prediction accuracy can reach 90% after the second college year and 99% after the third year. Among all the variables, cumulative grade points average (GPA) and nursing course GPA are the most influential factors for predicting graduation.

Conclusions: This study provides a potential mode of data-based tracking system for nursing students during their entire baccalaureate program. This tracking system can serve a large number of students automatically to provide customized evaluation on the dropout risk students and enhance the ability of a school or college to more strategically design school-based prevention and interventional services.

1. Introduction

Nursing comprises the largest workforce in health care. In 2018 it was estimated that 3,059,800 nurses were active in the workforce (Bureau of Labor Statistics, 2019). Despite this number, there is a national nursing shortage that does not appear to be lessening in the near future. It is estimated that, by 2028, the nation will experience a need for 3.4 million nurses to meet the rising demand in health care. This is a 12% increase over the current number of actively working nurses (Bureau of Labor Statistics). To address this challenge nursing schools are working diligently to enroll a sufficient number of students. Despite powerful efforts to maximize nursing school enrollment, schools and colleges of nursing are faced with high rates of attrition and low rates of completion. In addition to attrition rates influencing workforce numbers, regional (e.g., Higher Learning Commission) as well as specialized accrediting agencies, most notably the Commission on Collegiate

☆ This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
* Corresponding author.
E-mail addresses: LiHannaford@creighton.edu (L. Hannaford), xycheng@unomaha.edu (X. Cheng), MaryKunes-Connell@creighton.edu (M. Kunes-Connell).

https://doi.org/10.1016/j.nedt.2021.104784
Received 21 October 2019; Received in revised form 21 December 2020; Accepted 12 January 2021
Available online 26 January 2021

Nursing Education (CCNE) and the Accreditation Commission for Education in Nursing (ACEN), commonly use graduation rate as a metric to evaluate program effectiveness. CCNE requires that schools calculate the graduation rate for all programs and provide a written rationale should the completion rate fall below 70%. ACEN requires that a school or college sets its expected completion rates. In addition, many state boards of Nursing require a minimum pass rate before placing a school on probation.

Because of the high stakes ramifications of high attrition and low completion rates for schools, a number of studies have addressed attrition and completion rates in nursing programs. Early identification of at-risk students and the factors associated with graduation outcomes are the main foci for these studies. Jeffreys (2007) and Newton et al. (2007) suggested that early identification of at-risk students, variables impacting completion rates, and early intervention strategies were the common themes in nursing literature. Symes et al. (2005) identified a strong association between reading comprehension scores for the Nursing Entrance Test (NET) and retention to graduation. Using two cohorts of first-semester sophomore nursing students, Newton et al. (2007) suggested that both pre-nursing course scores and Test of Essential Academic Skills (TEAS) scores were useful predictors of early academic achievement. As cited by Jeffreys, in her 2007 work, researchers (Wharrad et al., 2003; Alexander and Brophy, 1997; Gallagher et al., 2001; Lin et al., 2003; Sadler, 2003) found that success in pre-nursing courses, as well as the first nursing course, played a significant role in predicting future success in the nursing program and/or licensure exam. Jeffreys's study was also consistent with the findings. The importance of performance in nursing courses for completion has also been an area of focus. Jeffreys et al. (2007) pointed out that students who withdrew or failed a nursing course at any point in the curriculum were at-risk. The findings from a study by Abele et al. (2013) supported the nursing literature that indicated certain nursing courses could predict the likelihood of success in completion of a nursing program. Though these studies are quite useful, a limitation to generalizing their results lies in their small sample sizes and inability to analyze large data sets from which to reliably identify potential risk factors and trend the data over time in order to determine the reliability of the results. Betts et al. (2017) surveyed 163 first year BSN students using the College Persistence Questionnaire. Their study suggested that variables such as time management worries, academic workload worry, and support system concerns ranked impacted potential for attrition. While studies such as these are useful in understanding at-risk variables, their limitation lies in a lack of a standardized method that could be used with large databases of students to develop a reliable predictive model. In addition, many of the studies were based on small sample sizes that limit their generalizability.

Various studies have identified factors related to academic success among nursing students in different scenarios using traditional methods. Machine learning, as a branch of artificial intelligence, is a collection of more sophisticated computing algorithms which learn from data and identify patterns for prediction and decision making. This technology has been shown to perform better in prediction tasks than traditional statistical methods (Breiman, 2001a; Donoho, 2017). Compared to the rule-based approaches commonly used in educational environments, machine learning techniques can identify at-risk students earlier and more accurately with a large amount of data. A few studies have been conducted in identifying at-risk students using machine learning algorithms in different educational scenarios. Lakkaraju et al. (2015) outlined a machine learning framework to identify high school students who were at risk of adverse academic outcomes. The data of two cohorts in two districts, altogether approximately 25,000 students, were analyzed using five machine learning techniques. The data covered demographic, academic performance, and behavioral attributes. The study showed that GPA at 8th grade was highly ranked across the majority of the approaches and the Random Forest model outperformed all the other models for both school districts. Other researchers studied the topic in online learning environment. Kotsiantis et al. (2003) collected demographic and performance records of 354 students and used six common machine learning algorithms to predict dropout students in a distance learning baccalaureate program. They concluded that the accuracy reached 63% in the initial predictions based only on demographic data and exceeded 83% before the middle of the academic period. The study also indicated that the Naïve Bayer and K-Nearest Neighbor had the best accuracy in the data sets. Another study for an online program by Lykourentzou et al. (2009) achieved a 75–85% overall student classification rate from the first section of the two courses, and reached a 97–100% rate in the final sections. In addition, it appeared that demographic characteristics were not found to aid in accurate dropout prediction. In higher education, Aulck et al. (2016) carried out a study to predict student dropout for baccalaureate program using 32,538 student data of demographic, pre-college entry information, and first academic term transcript records. They suggested that dropout could be accurately predicted with the three machine learning approaches employed even when predictions were based on a single term of academic transcript data. Logistic Regression outperformed the other two algorithms and reached the highest accuracy of 67%. However, machine learning technology, in the area of nursing education, has rarely been used as an analytical method to address retention. A review of the literature would indicate that the only study that has predicted the dropout behaviors of nursing students with a machine learning exercise was conducted in 2007, by Moseley and Mead (Moseley and Mead, 2008). Their study predicted whether a student would drop a nursing course before and during the course utilizing the demographic and academic performance of the student. The study achieved a 31% sensitivity at the start of the course and 84% sensitivity (94% accuracy) during the course. However, the dropout prediction of a nursing program was not conducted by student-level data analysis in the previous studies.

The purpose of this study is to determine the effectiveness of popular machine learning algorithms to build multiple models and predict possible dropout students at the beginning of each academic year starting from the first enrollment in a traditional baccalaureate nursing program. In addition, factors related to successful graduation are to be identified using several of the algorithms.

2. Objectives

This study attempted to identify adequate models to predict, early in a student's career, if an undergraduate nursing student will graduate within six college years. Therefore the following research questions were studied:

1. Can a machine learning model predict a nursing student's graduate status early in the student's academic career?
2. Which model is more effective and provides a more accurate rate?
3. What are the important predicting factors for nursing graduation?

3. Design

This study used a quantitative descriptive design. Approval was obtained from the Institutional Review Board of a medium-sized, private university in an urban Midwest city. All personal information was de-identified.

3.1. Participants

All records of full time, first time, and degree-seeking students who enrolled from 2004 through 2012 in a traditional 4-year baccalaureate nursing program were collected from the university's registrar databases. The final total number of qualified students for this study is 773, of which 245 students did not graduate within six years.

3.2. Data collection

The data contained student information from four categories: demographic background, high school information, college grade points average (GPA) scores, and graduation status. The demographic background included gender, marital status, ethnicity, religion, citizenship, residence, campus visit before enrollment, college dismissal, felony conviction, and first generation. The high school information included high school reported GPA, high school state, and years between high school graduation and current college enrollment. The college GPA category included all the term GPAs and cumulative GPAs by semester. The graduation status after six college years was categorized into two types: completed and not completed. All variables were collected from the University's database system.

3.3. Data preparation

Missing values occurred in 11 variables mostly in the areas of demographic background and high school information. Ethnicity had the largest missing value proportion at 34.7%, high school GPA had 23.9% missing values, and all other variables had less than 4% values missed. For categorical variables, all the missing values were assigned into a new category "not reported". The missing values for "years between high school graduation and current college enrollment" were replaced by the most common value of zero. For high school GPAs, missing values were imputed by multiple imputation (Van Buuren and Groothuis-Oudshoorn, 2011).

College GPAs were used to generate 63 variables for modeling. First, the courses graded as "pass" or "not pass" were excluded. Second, fall and winter terms were combined, and spring and summer terms were combined within each year in order to avoid too many missing values in winter and summer semesters. Thus, there were 12 terms in total for six college years. Third, term GPAs were separated by non-nursing and nursing courses, following previous studies (Abele et al., 2013; Jeffreys, 2007; Newton et al., 2007) which showed the necessity of differentiating the general courses and nursing courses. Eventually, variables of term GPA, cumulative GPA, nursing course GPA, non-nursing course GPA, and term credit hours for each of the 12 semesters were generated for each student. If a term GPA score or term credit hours did not exist for a certain term, it was then replaced with zero. Furthermore, based on the GPA scores, three new variables, number of total terms attended, number of terms with non-nursing courses attended, and number of terms with nursing courses attended, were created and included in the dataset.

Table 1 lists all the 77 variables used in the model. The response variable is binary indicating whether a student was graduated or not. The predictor variables were divided into three sections: demographic background, high school information, and college GPA variables.

3.4. Methods

Eight popular machine learning algorithms were chosen for model construction and comparison, including C5.0, random forest, xgboost, neural networks, support vector machine, Naïve Bayes, K-nearest neighbor, and logistic regression. C5.0, random forest, and xgboost are tree-based algorithms. C5.0 (Quinlan, 1993; Kuhn and Quinlan, 2018) is a decision tree that splits a population into branch-like sub-spaces such that values in each terminal node vote for the classification result. A visual example of C5.0 decision tree is shown in Fig. 1. Random forest (Breiman, 2001a, 2001b; Liaw and Wiener, 2002) is a combination of tree predictors such that each tree only depends on the observations from a bootstrap sample (Efron and Tibshirani, 1994) and variables from a random choice. Random forest optimizes the result by adapting the outcome of the majority vote of multiple random generalized trees. Xgboost (Chen and Guestrin, 2016) is also a tree-based ensemble algorithm but through a regularized gradient tree boosting approach

Table 1
Variables prepared for modeling.

Predictor variables (76 variables)			Response variable
Section 1	Section 2	Section 3	
Demographic background	High school	College GPA	
1. Gender 2. Marital status 3. Ethnicity 4. Religion 5. Citizenship 6. Residence 7. Campus visit before enrollment 8. College dismissal 9. Felony conviction 10. First generation	1. High school GPA 2. High school state 3. Years between high school graduation and current college enrollment	1. Term GPA (Term 1-12) 2. Cumulative GPA (Term 1-12) 3. Nursing term GPA (Term 1-12) 4. Non-nursing term GPA (Term 1-12) 5. Term credit hours (Term 1-12) 6. Number of terms attended 7. Number of terms with nursing courses attended 8. Number of terms with non-nursing courses attended	Graduation status (yes: 1/ no: 0)

(Friedman et al., 2000; Friedman, 2002), which produces an accurate prediction rule by combining rough and moderately inaccurate rules. Inspired by the nervous systems of human brain, neural networks connects the input and output through a sequence of hidden layers of neurons where the neuron connection weights are estimated via back-propagation (LeCun et al., 1990; Günther and Fritsch, 2010). Support vector machines (Boser et al., 1992) attempts to separate two classes of data using an optimal hyperplane that maximizes the margin between the two classes. Naïve Bayes is a generative probabilistic classifier that uses Bayes' theorem to predict the most probable class for a new observation from the given features (Rish, 2001). K-nearest neighbor (Cover and Hart, 1967) forms a majority vote between the K closest observations to a new observation. Logistic regression is one of the most common regression models to predict a binary outcome. It calculates the probability of an observation given the input features and assigns the observation to a discrete set of classes.

To achieve better accuracy for each method, we applied cross-validation (Friedman et al., 2001) to tune the parameters. The cross-validation approach splits the data into multiple folds, takes one-fold as validation set and the other folds as training set, iteratively trains the data with different training sets, and averages the prediction errors from validation sets. The hyperparameters giving lower cross validation error were selected.

In addition to the eight algorithms, a stacked ensemble method was introduced in the study. Stacking is an ensemble machine learning algorithm which combines the results of different predictive models to boost the accuracy and reduce the variance of prediction. After gaining the predicted graduation outcome for each student in the test dataset from the previous eight models, we stacked the eight outcomes by linear combination with a weight to each of the outcome, and then we calculated the result for each student using the formula below.

$$\widehat{Y_9} = I\left(\sum_{i=1}^{8} w_i \widehat{Y}_i > 0.5\right), w_i = \begin{cases} 0.18 & DT, RF, LR \\ 0.09 & Otherwise \end{cases}$$

where $\widehat{Y}_i \in \{0,1\}$ is the prediction of graduation status. $i = 1, 2, \cdots, 8$ indicates the eight algorithms, while $i = 9$ indicates the stacked ensemble method. w_i is the ensemble weight of the eight algorithms. Double weight was assigned to decision tree, random forest, and logistic regression because they were the top three algorithms that delivered the highest accuracies from various model designs, as shown later in

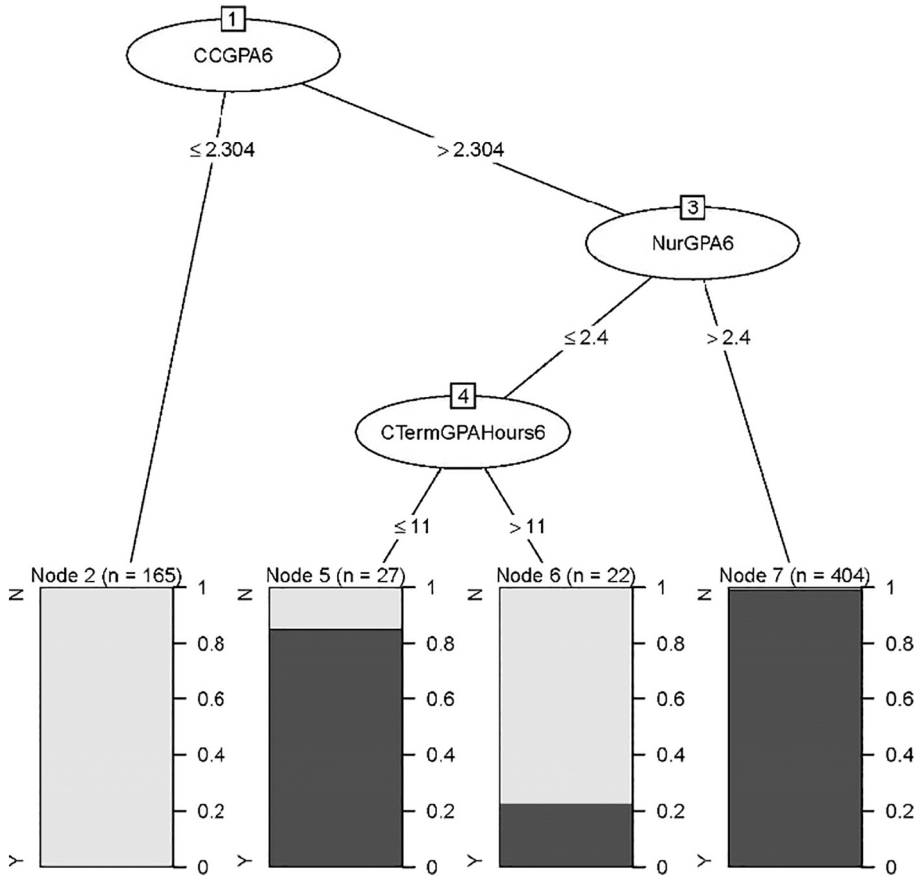

Fig. 1. Decision tree created by C5.0 method.

Table 3. Since the total weight is $\sum_{i=1}^{8} w_i = 1$, we used 0.18 for decision tree, random forest, and logistic regression models and 0.09 for the rest of the models. The final prediction of the stacked model, $\widehat{Y_9}$, was classified as a binary outcome based on the weighted average value.

3.5. Software

R language was employed to prepare and analyze the data.

3.6. Model design

First, the dataset was split into training and test sets. Models were built up on the training set and then validated on the test set. Validation on the test set would relieve the overfitting problem and provide an estimate of model accuracy for future data. The accuracy on the test set is usually used for model selection. Common ratio in machine learning 80/20 was used to split the data. Therefore, 618 out of 773 students were assigned into the training set and 155 into the test set. The nine machine learning models were trained on the training set with three combinations of variables at five different time points. Prediction results made on the same test set were compared in Section 4.

To identify at-risk students and provide necessary assistance along the study process before they withdraw from the program, five time points were chosen at the beginning of year 1, 2, 3, 4, and the time of graduation. 86% of the students in the data graduated in four years, but graduation within six years was also accepted. So in our model the time points were represented by year 0, 1, 2, 3, 5.

To examine the variable effectiveness, we designed three combinations of variable sections. First, all three variable sections were used. Second, demographic background and college GPAs were chosen as they give higher accuracy rates compared to the other two section combinations. Lastly, college GPA variables were employed alone for the prediction.

Table 2 displays all the design run for nine algorithms in the study. Fourteen scenarios were built for each algorithm through the combinations of different variable sections and time points. The model for year 0 with college GPA variables is not available due to non-existence of the data. The college GPA data was not available for the year 0 models in the other two combinations. Altogether, 126 models were built in the experiment.

4. Results

Over the nine-year entry period, the graduation rate varies between 63.5% and 74.5%, the median is 66.7%. The graduation rate has been trending higher from 64.4% of 2004 enrolled students to 74.5% of 2012 enrolled students. This section will show the performance of 126 models and compare the nine algorithms. In addition, we will make a suggestion of variables that have important impact on graduation at different time points.

Table 2
Model design.

Time point	Variable section		
	Demographic background + high school + college GPA	Demographic background + college GPA	College GPA
Year 0	X	X	–
Year 1	X	X	X
Year 2	X	X	X
Year 3	X	X	X
Year 5	X	X	X

4.1. Model accuracy

Overall accuracy criterion was used to measure the proportion of successfully predicted completers and non-completers out of the total number of students.

Table 3 displays the overall accuracies of all the models. If using stacked models as the reference, we find that the overall accuracy of stacked models increases from 74% to 100% as the participants grow from year 0 to full year 5 datasets regardless of variable sections. All the models using six-year data reach an overall accuracy greater than 98% for all the algorithms.

The comparisons of the stacked model overall accuracy by variable sections is exhibited in Fig. 2. When college GPA data is available, using GPA and term variables alone provides close or higher stacked overall accuracy rates compared to using the other two variable sections. The results are in line with the discovery of Aulck et al. (2016) that the outcome can be accurately predicted even when the prediction is based on a single term of academic transcript data. The importance of this finding is that the dimension of sample space needed for accurate outcome prediction can be significantly reduced. Graduation outcome can be correctly predicted for greater than 80% of the students with just one year of college GPA data. By the time the second academic year is completed, the prediction accuracy rate reaches above 90%. Demographic background and high school information does not appear to have a significant impact on the prediction of graduation outcome.

4.2. Algorithm performance

It is a common practice to use the Area Under Curve (AUC) of Receiver Operating Characteristic (ROC) to evaluate the performance of machine learning algorithms for binary classification. Better model will have a higher AUC value.

Table 4 displays the AUC values of each algorithm for all the models. Overall, random forest, logistic regression, and decision tree present better performance among all the individual models compared to the other five algorithms. As stated earlier, more weight was given to these three algorithms when stacked models were carried out. We can see that stacked models deliver the highest AUC scores for approximately 60% of the 14 models and rank top 1 or 2 performed models for close to 80% of all the models. Individually, random forest outperforms other seven algorithms by providing the highest AUC scores for about 60% of the models.

4.3. Sensitivity

While overall accuracy is important for the prediction, the capability of correctly identifying potential dropout students is more crucial for the program as it will enable the college to provide suitable assistance to the at-risk students as early as possible. Sensitivity criterion can be used to measure the efficiency with which a machine learning technique can accurately identify dropout students (Lykourentzou et al., 2009). It is the proportion of correctly predicted non-completers out of the actual number of non-completers, i.e., $1-\beta$, where β is the type II error. Higher value in sensitivity means more students who cannot graduate can be successfully identified. It is important to check this criterion because compared to type I error which identifies graduated students as non-completers, lower sensitivity has worse influence, as it will ignore the students who need help at an early stage.

Table 5 presents the sensitivity values for all stacked models as they outperformed other models for most of the 14 scenarios. Stacked models were able to accurately identify 21% of the non-completers at the beginning of the study regardless the variable section used. The sensitivity increases to approximately 50% for all the models when the first year of study is completed. At the beginning of the third year, around 70%–80% of the non-completers can be correctly discovered, and by the end of the third year, 98% of the dropout students can be identified.

Table 3
Overall accuracies by algorithm, variable section, and time point.

Algorithm	Year 0			Year 1			Year 2			Year 3			Year 5		
	Demographic Background + High School	Demographic Background	College GPA	Demographic Background + High School + College GPA	Demographic Background + College GPA	College GPA	Demographic Background + High School + College GPA	Demographic Background + College GPA	College GPA	Demographic Background + High School + College GPA	Demographic Background + College GPA	College GPA	Demographic Background + High School + College GPA	Demographic Background + College GPA	College GPA
Naïve Bayes Classifier	45.2%	36.8%	-	74.2%	74.8%	78.1%	85.2%	86.5%	91.0%	93.5%	94.8%	96.8%	98.1%	99.4%	99.4%
Decision Tree	73.5%	68.4%	-	81.3%	81.3%	81.3%	89.0%	91.0%	92.3%	99.4%	99.4%	98.7%	98.7%	98.7%	98.7%
Random Forest	72.9%	73.5%	-	81.9%	81.9%	79.4%	91.6%	91.6%	93.5%	99.4%	99.4%	98.7%	100.0%	100.0%	100.0%
Boosting	69.7%	69.7%	-	80.6%	78.1%	80.0%	93.5%	93.5%	92.9%	98.1%	98.1%	98.1%	100.0%	100.0%	100.0%
K-Nearest Neighbor	64.5%	67.1%	-	74.2%	74.2%	72.3%	88.4%	87.7%	92.3%	96.1%	96.8%	97.4%	100.0%	100.0%	100.0%
Logistic Regression	71.0%	69.7%	-	80.6%	81.9%	81.3%	90.3%	90.3%	92.3%	99.4%	98.7%	98.7%	100.0%	100.0%	100.0%
Neural Network	65.8%	65.8%	-	74.8%	74.2%	81.3%	86.5%	87.7%	92.3%	92.3%	97.4%	99.4%	99.4%	99.4%	99.4%
Support Vector Machines	71.6%	69.7%	-	79.4%	78.1%	80.6%	89.7%	92.3%	89.7%	98.1%	98.1%	96.8%	100.0%	100.0%	99.4%
Stacked Model	74.2%	73.5%	-	81.9%	81.3%	81.3%	91.0%	91.6%	92.9%	99.4%	99.4%	99.4%	100.0%	100.0%	100.0%

*The highest values of the models in a specific category are highlighted in blue.

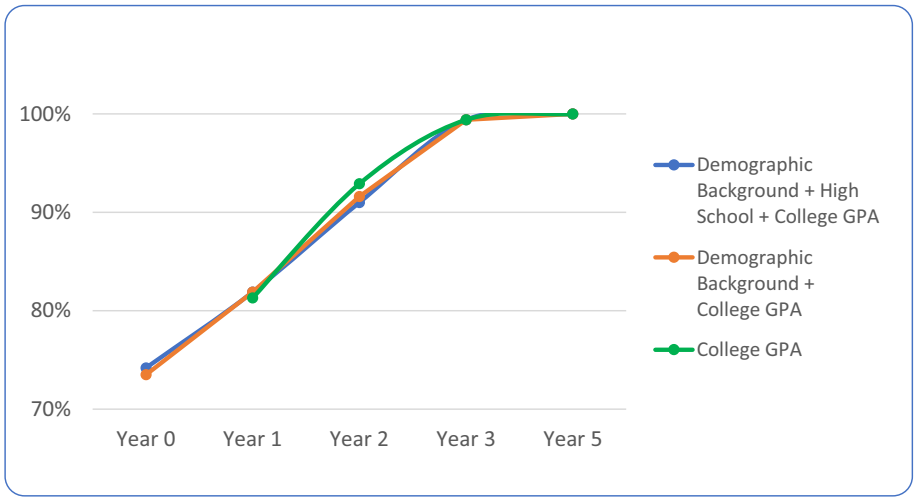

Fig. 2. Stacked model overall accuracy.

4.4. Variable importance

When machine learning approaches are conducted, predictor importance can be revealed from tree algorithms. The importance value is measured by how frequently the predictor is used and how much purity the predictor can contribute to the classification when constructing the tree. The top five predictors were identified by the three tree-based algorithms: decision tree (DT), boosting (BT), and random forest (RF), as shown in Table 6.

The prediction of a student's graduation prominently relies on the residence type for year 0 models. It was found by logistic regression that the graduation rate was significantly higher for the students who lived in a dorm compared to those who lived off campus. Starting from year 1, all variables regarding to college GPA and other college academic performance have positive impact on graduation. The cumulative GPA for the second term plays a key role for both year 1 and year 2 models. For year 3 models, cumulative GPA and nursing GPA for term 6 appear to be the crucial factors. Finally, the number of nursing terms attended and nursing GPA for term 8 are notably related to the completion outcome across the three algorithms. Overall, cumulative GPA score plays an important role for the graduation prediction. In addition, nursing GPA score starts showing the significance after two years of study.

5. Discussion

Nine machine learning algorithms in 14 scenarios of variable combinations at five time points were investigated to determine which model might be the best predictor of at-risk students in a four-year nursing baccalaureate program. Compared to other relevant research in nursing program graduation, this study focused more on improving the dropout prediction for each individual student and making prediction at an early stage.

This study demonstrates how machine learning methods can enable nursing educators to predict potential at-risk students on a yearly basis and identify them in a timely manner. With demographic and background information, high school records, and college GPA scores, graduation outcomes could be more accurately predicted for above 70% of first enrolled students and for close to 100% of students who completed their third year of study.

Despite the study's findings regarding the effectiveness of machine learning methods to predict potential at-risk, this investigation is not without its limitations. The following limitations need to be addressed in future studies.

5.1. Data quality of influential factors

For year 0 models, the important variables high school GPA and ethnicity had 24% and 35% of missing values. Although the missing values in high school GPA were imputed and in ethnicity were categorized separately, the impact of missing values was not studied in our work.

5.2. Asymmetric response variable

The outcome measure (whether the student graduated or not) can be asymmetric, as not only the dropout students, but also some of the graduated students, were actually at risk in obtaining the program degree, and would need early intervention. Our models did classify some graduated students into the dropout category for their similar behaviors to the dropout students, but to what extent they needed the intervention was not studied.

5.3. Low sensitivity at the early stage

Even though the overall accuracy rates of stacked models were above 70% and 80%, and the specificity rates (proportion of actual dropout students within the predicted dropout students) were above 83% and 85% for year 0 and year 1 models, only 20% and 50% of dropout students could be correctly identified for these models. Future research needs to focus on how to improve the sensitivity of the models and enhance the effectiveness of predicting non-completed students at the early stage.

5.4. Generalization of the study's findings

The property of the study's data limits generalization of the study's findings as the sample consisted of only traditional four-year nursing students in a private university. It would be more beneficial to replicate the study in other four-year programs, second-degree programs, associate degree programs, as well as nursing programs in public universities and community colleges.

6. Conclusion

Among the eight commonly used machine learning algorithms, random forest demonstrated improved performance compared to the other individual algorithms by delivering the highest AUC scores among approximately 60% of the scenarios. Regarding to the overall prediction

Table 4
AUC by algorithm, variable section, and time point.

Algorithm	Year 0			Year 1			Year 2			Year 3			Year 5		
	Demographic Background + High School	Demographic Background	College GPA	Demographic Background + High School + College GPA	Demographic Background + College GPA	College GPA	Demographic Background + High School + College GPA	Demographic Background + College GPA	College GPA	Demographic Background + High School + College GPA	Demographic Background + College GPA	College GPA	Demographic Background + High School + College GPA	Demographic Background + College GPA	College GPA
Naïve Bayes Classifier	0.557	0.519	-	0.715	0.726	0.657	0.829	0.844	0.860	0.948	0.957	0.954	0.986	0.995	0.990
Decision Tree	0.584	0.559	-	0.721	0.721	0.721	0.852	0.871	0.881	0.990	0.990	0.985	0.979	0.979	0.979
Random Forest	0.568	0.579	-	0.731	0.731	0.713	0.870	0.870	0.902	0.990	0.990	0.985	1.000	1.000	1.000
Boosting	0.545	0.539	-	0.716	0.698	0.706	0.907	0.902	0.891	0.980	0.980	0.980	1.000	1.000	1.000
K-Nearest Neighbor	0.576	0.601	-	0.675	0.675	0.673	0.830	0.831	0.892	0.949	0.954	0.970	1.000	1.000	1.000
Logistic Regression	0.577	0.568	-	0.728	0.726	0.727	0.861	0.861	0.881	0.990	0.985	0.985	1.000	1.000	1.000
Neural Network	0.534	0.540	-	0.674	0.669	0.727	0.816	0.848	0.886	0.898	0.958	0.990	0.990	0.990	0.990
Support Vector Machines	0.588	0.579	-	0.718	0.703	0.716	0.862	0.898	0.868	0.974	0.980	0.971	1.000	1.000	0.990
Stacked Model	0.595	0.590		0.731	0.727	0.721	0.860	0.870	0.891	0.990	0.990	0.990	1.000	1.000	1.000

*The highest values of the models in a specific category are highlighted in blue.

Table 5
Sensitivity and confusion matrix of stacked models on the test set.

Variable section		Year 0					Year 1						Year 2						Year 3						Year 5					
		Demographic background + high school		Demographic background		College GPA	Demographic background + high school + college GPA		Demographic background + college GPA		College GPA		Demographic background + high school + college GPA		Demographic background + college GPA		College GPA		Demographic background + high school + college GPA		Demographic background + college GPA		College GPA		Demographic background + high school + college GPA		Demographic background + college GPA		College GPA	
Sensitivity		0.208		0.208		–	0.500		0.500		0.479		0.729		0.750		0.792		0.979		0.979		0.979		1.000		1.000		1.000	
		Observed		Observed			Observed		Observed		Observed		Observed		Observed		Observed		Observed		Observed		Observed		Observed		Observed		Observed	
Confusion matrix[a]	Predicted	TRUE	FALSE	TRUE	FALSE		TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE
	TRUE	10	2	10	3		24	4	24	5	23	4	35	1	36	1	38	1	47	0	47	0	47	0	48	0	48	0	48	0
	FALSE	38	105	38	104		24	103	24	102	25	103	13	106	12	106	10	106	1	107	1	107	1	107	0	107	0	107	0	107

[a] In confusion matrix, TRUE represents for dropout students, and FALSE represents for students graduated.

Table 6
Variable importance.

	Demographics + High School + College GPA			Demographics + College GPA			College GPA		
	DT	BT	RF	DT	BT	RF	DT	BT	RF
Year 0	ResidenceDorm	HighSchoolReportedGPA	HighSchoolReportedGPA	ResidenceDorm	ResidenceDorm	ResidenceDorm			
		ResidenceDorm	ResidenceDorm		EthnicityWhiteNonHispanic	ResidenceOffCampus			
		EthnicityWhiteNonHispanic	ResidenceOffCampus		ReligionCatholic	EthnicityWhiteNonHispanic			
		ReligionCatholic	EthnicityWhiteNonHispanic		EthnicityAsianPacificIslander	ReligionCatholic			
		HighSchoolStateMN	StateNotReported		MaritalStatusMarried	EthnicityAsianPacificIslander			
Year 1	CumulativeGPA2	CumulativeGPA2	CumulativeGPA2	CumulativeGPA2	CumulativeGPA2	CumulativeGPA2	CumulativeGPA2	CumulativeGPA2	CumulativeGPA2
		HighSchoolReportedGPA	TermGPA2		TermGPA2	NonNursingTermGPA2		TermGPA2	NonNursingTermGPA2
		TermGPA2	NonNursingTermGPA2		TermGPA1	TermGPA2		TermGPA1	TermGPA2
		NonNursingTermGPA2	NonNursingTermGPA1		MaritalStatusNotMarried	TermGPA1		TermCreditHours1	CumulativeGPA1
		TermCreditHours1	CumulativeGPA1		TermCreditHours2	CumulativeGPA1		NursingGPA2	NonNursingTermGPA1
Year 2	CumulativeGPA4	CumulativeGPA4	CumulativeGPA4	CumulativeGPA4	CumulativeGPA4	CumulativeGPA4	CumulativeGPA4	CumulativeGPA4	CumulativeGPA4
	NursingTermGPA4	NursingTermGPA4	TermGPA4	NursingGPA4	NursingGPA4	NursingGPA4	NursingGPA4	TermGPA4	NursingGPA4
	NonNursingTermGPA1	TermGPA3	NursingGPA4	NonNursingTermGPA1	TermGPA3	TermGPA4	TermGPA3	NursingGPA4	TermGPA4
	TermGPA3	TermCreditHours1	TermCreditHours4	TermGPA3	TermCreditHours1	CumulativeGPA3	TermCreditHours1	TermCreditHours1	NonNursingTermGPA4
	EthnicityAsianPacificIslander	TermGPA1	TermGPA3	EthnicityAsianPacificIslander	TermGPA1	NonNursingTermGPA4	NonNursingTermGPA3	CumulativeGPA2	TermGPA3
Year 3	CumulativeGPA6	NursingGPA6	TermGPA5	CumulativeGPA6	NursingGPA6	NursingGPA6	CumulativeGPA6	NursingGPA6	NursingGPA6
	NursingGPA6	CumulativeGPA6	NursingGPA6	NursingGPA6	CumulativeGPA6	CumulativeGPA6	NursingGPA6	CumulativeGPA6	CumulativeGPA5
	NonNursingTermsAttended	TermGPA5	TermGPA6	NonNursingTermsAttended	TermGPA5	TermGPA5	NonNursingTermsAttended	TermGPA5	NursingGPA5
	NonNursingTermGPA6	TermGPA6	CumulativeGPA6	NonNursingTermGPA6	NursingGPA4	NursingGPA5	NonNursingTermGPA6	NursingGPA4	CumulativeGPA6
	TermGPA1	TermGPA3	NursingGPA5	TermGPA1	TermGPA6	TermsAttended	TermGPA1	TermGPA6	TermGPA5
Year 5	NursingGPA8	NurTermsAttended	NursingGPA8	NursingGPA8	NurTermsAttended	TermCreditHours8	NursingGPA8	NursingTermsAttended	NursingTermsAttended
	CumulativeGPA10	CumulativeGPA8	TermCreditHours8	CumulativeGPA10	CumulativeGPA8	TermGPA8	CumulativeGPA10	CumulativeGPA8	CumulativeGPA8
		CumulativeGPA7	NurTermsAttended		CumulativeGPA7	NurTermsAttended		CumulativeGPA7	TermCreditHours8
		TermCreditHours6	TermGPA8		TermCreditHours6	CumulativeGPA8		TermCreditHours6	TermGPA8
		NursingGPA5	TermsAttended		NursingGPA5	TermsAttended		NursingGPA5	NursingGPA8

Importance Level:

accuracy, the stacked models proposed by the weighted average of the result from eight algorithms outperformed all individual algorithms.

This study revealed that college GPA, nursing course grades and credit taken consisted the most important variables for predicting nursing graduation. Demographic and most of the high school information did not appear to have much impact on the prediction results. At the start of the college, the residence type and high school GPA were crucial factors. However, they were quickly replaced when the college GPA was added to the model. After the regular four years of study, the number of nursing terms attended played a more important role in the graduation prediction.

This study provides a potential mode of data-based tracking system for nursing students during their entire baccalaureate program. This tracking system provides customized evaluation on the dropout risk for each student using their own background and academic performance information. Increasing the specificity and accuracy of identifying at-risk variables will enhance the ability of a school or college to more strategically design school-based prevention and interventional services as well as individualized student plans to facilitate their success in a nursing program. In addition, the tracking system can serve a large number of students automatically, which will help the university to retain students at a low cost of labor, time, and financial expenditure. While the study was carried out in a nursing baccalaureate environment, the methodology has the capability to be applied to any other similar educational scenarios.

CRediT authorship contribution statement

Li Hannaford: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Visualization, Project administration.

Xiaoyue Cheng: Methodology, Software, Formal analysis, Investigation, Writing - Review & Editing, Visualization.

Mary Kunes-Connell: Conceptualization, Writing - Review & Editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abele, C., Penprase, B., Ternes, R., 2013. A closer look at academic probation and attrition: what courses are predictive of nursing student success? Nurse Educ. Today 33, 258–261.

Alexander, J.E., Brophy, G.H., 1997. A five-year study of graduates' performance on NCLEX-RN. J. Nurs. Educ. 36, 443–445.

Aulck, L., Velagapudi, N., Blumenstock, J., West, J., 2016. Predicting Student Dropout in Higher Education (Paper presented at 2016 ICML Workshop on #Data4Good: Machine learning in social good applications, New York, NY).

Betts, K.J., Shirley, J.A., Kennedy, R., 2017. Identifying academic and social risk factors of baccalaureate nursing students using the college persistence questionnaire. Journal of Education and Practice 8 (ISSN 2222-288x at www.iiste.org).

Boser, B., Guyon, I., Vapnik, V., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. ACM, pp. 144–152.

Breiman, L., 2001a. Random forests. Machine Learning 45, 5–32.

Breiman, L., 2001b. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Statistical science 16 (3), 199–231.

Bureau of Labor Statistics, U.S. Department of Labor, 2019. Occupational Outlook Handbook. Registered Nurses, on the Internet at. https://www.bls.gov/ooh/healthcare/registered-nurses.htm.

Chen, T., Guestrin, C., 2016, August. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 785–794.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13 (1), 21–27.

Donoho, D., 2017. 50 years of data science. J. Comput. Graph. Stat. 26 (4), 745–766.

Efron, B., Tibshirani, R., 1994. An Introduction to the Bootstrap. CRC press.

Friedman, J., 2002. Stochastic gradient boosting. Computational statistics & data analysis 38 (4), 367–378.

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann. Stat. 28 (2), 337–407.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning (Vol. 1, No. 10). Springer series in statistics, New York.

Gallagher, P., Bomba, C., Crane, L., 2001. Using an admissions exam to predict student success in an ADN program. Nurse Educ. 26(3), 132–135.

Günther, F., Fritsch, S., 2010. neuralnet: training of neural networks. The R journal 2 (1), 30–38.

Jeffreys, M., 2007. Tracking students through program entry, progression, graduation, and licensure: assessing undergraduate nursing student retention and success. Nurse Educ. Today 27, 406–419.

Kotsiantis, S.B., Pierrakeas, C.J., Pintelas, P.E., 2003. Prevent student dropout in distance learning using machine learning techniques. In: Palade, V., Howlett, R.J., Jain, L.C. (Eds.), Knowledge-based Intelligent Information and Engineering Systems, KES 2003. Lecture Notes in Computer Science, vol. 2774. Springer, Berlin, Heidelberg, pp. 267–274.

Kuhn, M., Quinlan, R., 2018. C50: C5.0 decision trees and rule-based models. R package version 0.1.2. https://CRAN.R-project.org/package=C50.

Lakkaraju, H., Miller, D., Aguiar, E., Bhanpuri, N., Addison, K., Shan, C., Ghani, R., 2015. A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes. Paper presented at 21th ACM SIGKDD International Conference, Sydney, Australia. https://doi.org/10.1145/2783258.2788620.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L., 1990. Handwritten digit recognition with a back-propagation network. In: Advances in Neural Information Processing Systems, pp. 396–404.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22.

Lin, R.S.J., Fung, B.K.P., Hsiao, J.K., Lo, H.F., 2003. Relationship between academic scores and performance on national qualified examination for registered professional nurses (NQEX-RPN). Nurse Educ. Today 23, 492–497.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., Loumos, V., 2009. Dropout prediction in e-learning courses through the combination of machine learning techniques. Computer & Education 53, 950–965.

Moseley, L.G., Mead, D.M., 2008. Predicting who will drop out of nursing courses: a machine learning exercise. Nurse Educ. Today 28, 469–475.

Newton, S., Smith, L., Moore, G., Magnan, M., 2007. Predicting early academic achievement in a baccalaureate nursing program. J. Prof. Nurs. 23 (3), 144–149.

Quinlan, J., 1993. C4. 5: Programs for Machine Learning. Morgan Kaufmann Publishers.

Rish, I., 2001. An empirical study of the naive Bayes classifier. In: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, Vol. 3, No. 22, pp. 41–46.

Sadler, J., 2003. Effectiveness of student admission essays in identifying attrition. Nurse Educ. Today 23, 620–627.

Symes, L., Tart, K., Travis, L., 2005. An evaluation of the nursing success program. Nursing Educator 30 (5), 217–220.

Van Buuren, S., Groothuis-Oudshoorn, K., 2011. mice: multivariate imputation by chained equations in R. J. Stat. Softw. 45 (3), 1–67.

Wharrad, H.J., Chapple, M.A., Price, N., 2003. Predictors of academic success in a Bachelor of Nursing course. Nurse Educ. Today 23, 246–254.