The Application of Adaptive Minimum Match k-Nearest Neighbors to Identify At-Risk Students in Health Professions Education

Anshul Kumar, PhD; Taylor DiJohnson, BA; Roger A. Edwards, ScD; Lisa Walker, MPAS, PA-C

**Introduction** When learners fail to reach milestones, educators often wonder if any warning signs could have allowed them to intervene sooner. Machine learning can predict which students are at risk for failing a high-stakes certification examination. If predictions can be made well before the examination, educators can meaningfully intervene before students take the examination to reduce their chances of failing.

**Methods** The authors used already-collected, first-year student assessment data from 5 cohorts in a single Master of Physician Assistant Studies program to implement an “adaptive minimum match” version of the k-nearest neighbors algorithm using changing numbers of neighbors to predict each student’s future examination scores on the Physician Assistant National Certifying Exam (PANCE). Validation occurred in 2 ways by using leave-one-out cross-validation (LOOCV) and by evaluating predictions in a new cohort.

**Results** “Adaptive minimum match” version of the k-nearest neighbors algorithm achieved an accuracy of 93% in LOOCV. “Adaptive minimum match” version of the k-nearest neighbors algorithm generates a predicted PANCE score for each student one year before they take the examination. Students are classified into extra support, optional extra support, or no extra support categories. Then, one year remains to provide appropriate support to each category of student.

**Discussion** Predictive analytics can identify at-risk students who might need additional support or remediation before high-stakes certification examinations. Educators can use the included methods and code to generate predicted test outcomes for students. The authors recommend that educators use predictive modeling responsibly and transparently, as one of many tools used to support students. More research is needed to test alternative machine learning methods across a variety of educational programs.
Predictive Analytics in Education

A typical approach to educational analytics is to use student performance or behavior data at intermediate stages to predict a final outcome of interest. For example, student results on homework assignments or quizzes and student activity records in online learning management systems during a semester-long course can be used to make guesses about final examination performance. Predicting subsequent outcomes based on earlier performance has been demonstrated with students studying many topics, including informatics,\(^{5}\) human computer interaction,\(^{6}\) computer hardware,\(^{7}\) and mathematics.\(^{7}\) Liz-Domínguez et al\(^{8}\) refer to this method as grade prediction. The same approach has also been used to predict which students will withdraw from an educational program.\(^{9,10}\) These studies train and evaluate the utility of machine learning models. Such machine learning models can be used on future groups of students to create early warning systems that identify at-risk students and intervene, as described by Arnold and Pistilli.\(^{11}\)

Within health professions specifically, commentaries on machine learning abound,\(^{12-14}\) whereas empirical studies that apply machine learning are less common. Black et al\(^{15}\) take a similar approach to ours within the PA education context (which we built on by using additional predictive models, validating our approach has also been used to predict which students will withdraw from an educational program).\(^{9,10}\) These studies train and evaluate the utility of machine learning models. Such machine learning models can be used on future groups of students to create early warning systems that identify at-risk students and intervene, as described by Arnold and Pistilli.\(^{11}\)

This research was approved by the Mass General Brigham Institutional Review Board in 2020 (Protocol No. 2020P000514).

Our study uses student performance data from a Master of PA Studies program to create a predictive model to identify students at potential risk for failing the PANCE. Background information about this PA program and data are available in the supplemental Study Appendix (Supplementary Digital Content 1, http://links.lww.com/PAEA/A62). Table 1 summarizes the process of data collection, analysis, and application of results as we use it in practice. We used R version 3.6.3 (R Foundation for Statistical Computing) for all analyses. We also developed an open-source R package, AdaptiveLearnalytics,\(^{3}\) to perform key portions of our analysis. Our entire analysis is provided in the Study Appendix (Supplementary Digital Content 1, http://links.lww.com/PAEA/A62).

**Data Description and Preprocessing**

We used deidentified student data from 5 successive cohorts, starting in 2015—a total of 224 students—from a single PA program. The data were organized such that each row is a student and each column is a variable (containing a measured grade or other characteristic). We used the individual Readiness Assurance Test (iRAT) quiz, final examination, final course grade, and the Physician Assistant Clinical Knowledge Rating and Assessment Tool (PACKRAT) scores from students’ first year in the PA program and undergraduate grade point average (GPA) data. Although regular assessments were conducted throughout the second year of the PA program, we did not use these data or any demographic data. We excluded second-year data so that an entire year is available to provide additional support to students who are at risk for failing PANCE.

Our dependent variable is the PANCE score from each student’s first attempt at this examination after the student completes the 25-month PA program. PANCE scores can range from 200 to 800 points, and students must score 350 or higher to pass. We used student performance variables in the dataset, described below, as independent variables to predict students’ PANCE scores.

This dataset, collected as part of this PA program’s mostly team-based learning (TBL) curriculum, yielded more information about each student than we would expect from a traditional (non-TBL) curriculum. For each TBL course, we calculated the mean of each student’s iRAT scores and eliminated the individual iRAT score variables. For example, if students completed 7 iRAT quizzes during course A, we made a new variable with the mean of those 7 iRAT quiz scores. We then eliminated the 7 individual iRAT quiz scores from our dataset.

These preprocessing steps yielded the following independent variables for each student for 18 first-year courses: iRAT mean (when applicable), final examination score, and final grade. We also included overall undergraduate GPA, undergraduate science GPA, and first-year PACKRAT score. As described in the Study Appendix (Supplementary Digital Content 1, http://links.lww.com/PAEA/A62), we tried 2 different preparations (A and B) of these independent variables to determine which one yielded better results for each type of predictive model.

After completing the manual preprocessing of variables, we further narrowed down our independent variables by calculating the Pearson correlation coefficient of each independent variable with the PANCE score. Any variable correlated with PANCE below a selected threshold was eliminated, as shown in the Study Appendix (Supplementary Digital Content 1, http://links.lww.com/PAEA/A62). Before training any predictive models, we standardized all independent variables.

**METHODS**

This research was approved by the Mass General Brigham Institutional Review Board in 2020 (Protocol No. 2020P000514).

Our study uses student performance data from a Master of PA Studies program to create a predictive model to identify students at potential risk for failing the PANCE. Background information about this PA program and data are available in the supplemental Study Appendix (Supplementary Digital Content 1, http://links.lww.com/PAEA/A62). Table 1 summarizes the process of data collection, analysis, and application of results as we use it in practice. We used R version 3.6.3 (R Foundation for Statistical Computing) for all analyses. We also developed an open-source R package, AdaptiveLearnalytics,\(^{3}\) to perform key portions of our analysis. Our entire analysis is provided in the Study Appendix (Supplementary Digital Content 1, http://links.lww.com/PAEA/A62).

**Table 1. Entire 2-Year Analytics Process From Program Perspective**

| Time       | Prematriculation | Year 1 | Year 2 | Year 2 |
|------------|------------------|--------|--------|--------|
| Process    |                  |        |        |        |
| step       |                  |        |        |        |
| GPA        | Gather data:     | Gather data: iRAT quizzes, final    | Run machine learning  | Provide additional support All students take |
|           | undergraduate GPA | examinations, final course grades, | model to identify at-risk | students to predicted at-risk |
|           |                  | PACKRAT score |      |        |
|           |                  |        |        |        |

GPA, Grade Point Average; iRAT, Individual Readiness Assurance Test; PACKRAT, Physician Assistant Clinical Knowledge Rating and Assessment Tool; PANCE, Physician Assistant National Certifying Exam.
KNN procedures have been used in different contexts.27-29 Standard KNN are in our Study Appendix (Supplementary Digital Content 1, http://links.lww.com/PAEA/A62). As shown in our results, none of these techniques allowed us to create a predictive model that would achieve our educational goals of prioritizing the detection of students who might fail the PANCE while minimizing incorrect (false-positive) predictions.

For the modified version of KNN—AMMKNN—that we developed, the following steps summarize how we predicted each student’s PANCE score: (1) We went one by one through each of our current-year students; (2) for each current student, the computer accesses a data spreadsheet of our alums (previous graduates) to determine which 20 alums are most similar to this current student (using the independent variables to determine how similar or different each alum is from our current student); (3) we then created a series of averages based on these 20 alums’ PANCE scores and took the minimum of these averages (the most conservative estimate) to make a guess about our current student’s PANCE score; (4) we then repeated this procedure for each of our remaining current-year students so that we had a predicted PANCE score for each student who had not yet taken the PANCE. Like most standard analytics approaches, AMMKNN uses patterns and results from alums to make guesses about current students. The difference in using AMMKNN versus KNN is that AMMKNN attempts to calculate a “worst-case scenario” estimate for each student.

We can run this procedure for our current students each year, once they finish the first year of our PA program, to help us decide whom to recommend for our extra support program. The “adaptive minimum match” modification to KNN makes our approach more useful: very likely to fail (predicted PANCE score < 350), moderate risk of failing (predicted PANCE score 350–375), and low risk of failing (predicted PANCE score > 375). The moderate risk category ensured we could identify more at-risk students, not only those who might fail. This additional category supported our goal of using predictive analytics to help students advance toward examination with the best preparation.

Before settling on a single predictive technique, we used a number of standard approaches, such as random forest (RF), support vector machine (SVM), and standard k-nearest neighbors (KNN). These attempts are presented in the Study Appendix (Supplementary Digital Content 1, http://links.lww.com/PAEA/A62). As shown in our results, none of these techniques allowed us to create a predictive model that would achieve our educational goals of prioritizing the detection of students who might fail the PANCE while minimizing incorrect (false-positive) predictions.

For each predictive model, we created a standard 2-by-2 confusion matrix to show each student’s predicted value (from when they were the testing student) and their actual PANCE result. From the confusion matrices for each model, we calculated the following standard metrics: true positives, false positives, false negatives, accuracy, sensitivity, and specificity. Definitions of these metrics are in our Study Appendix (Supplementary Digital Content 1, http://links.lww.com/PAEA/A62).

To make our predictions more useful, we further aggregated them into 3 groups: predicted to fail (PANCE < 350), at risk of failing (350–375), and likely to pass (>375). This approach is analogous to the “traffic signal” strategy.11 We argue that sorting students into these 3 groups is most useful for our goals of optimizing potential remediation well in advance of the examination. In addition to traditional 2-by-2 confusion matrix data in Table 2, we present 3-by-3 confusion matrices in Figure 1 that show all 3 groups. We then discuss the practical application of the 3-group framework within our educational context.

RESULTS

Table 2 summarizes the LOOCV results of multiple predictive models based on standard 2-by-2 confusion matrices. AMMKNN has the highest accuracy with both variable preparations (0.93 with preparation B and 0.91 with preparation A). Because our priority as educators is to minimize false negatives (students who “fall through the cracks,” meaning they need our help but we fail to detect this need), we wanted to maximize sensitivity while keeping false positives (“unnecessary support” students who are flagged by the model but do not truly need remediation) within reasonable limits. The model with the highest sensitivity was RF(A), which predicted that 33 students would fail, 10 of which were correct predictions and 23 were incorrect “unnecessary support” predictions. Providing remediation to 33 students when only 10 of them (less than one-third) need it is not reasonable, we argue. The next-highest sensitivity was 0.69, shared by AMMKNN(B) and SVM(B). AMMKNN also had higher accuracy and specificity. AMMKNN(B) predicted that 18 students would fail, with 9 of these predictions being correct and 9 being incorrect. In this scenario, half of all students flagged for extra support would truly require it. SVM(B) predicted that 27 students would fail, with 9 of these predictions being correct and 18 being incorrect.
To break down the predictions more clearly and define how we can use them in practice, Figure 1 shows 3-by-3 confusion matrices for selected models. The accuracy for AMMKNN(B) was calculated as the number of correct predictions divided by the number of total students: \( \frac{9 + 3 + 124}{181} = 0.75 \). Although this accuracy was lower than the accuracy of the same model when a 2-by-2 matrix was used (0.93), we argue that the 3-by-3 version—which views results as a spectrum of 3 classification categories—is more useful when considering the use of the model in practice.

We can build a framework based on the 3-by-3 AMMKNN(B) results: 18 students are predicted to fail the PANCE. Of those 18 predictions, 9 would be completely correct, 1 would be justifiable (because it would be acceptable to remediate a student who would otherwise almost fail), and 8 would be “unnecessary support” students who receive remediation although it would not have been needed. The model would identify 26 students as “at risk.” Of these 26 predictions, 5 would be useful (2 students who truly fail and 3 students “at risk”) and 21 would be “unnecessary support” students. The model would identify 137 students as likely to pass. Of these 137 predictions, 124 would be “predictable no support” students who pass without risk, 11 would pass with an “at-risk” score, and 2 would “fall through the cracks” because we failed to identify them as needing remediation. The results for AMMKNN(A) are similar, whereas SVM(B) and standard KNN(B) each have 16 false positives, which is likely too many “unnecessary support” students to reasonably remediate. Our recommended use of these results is to remediate many students with predicted “fail” scores, to consider remediating predicted “at-risk” students on a student-by-student basis, and not to remediate students with predicted “pass” scores (unless extenuating circumstances or other information suggest otherwise).

After completing the validation and inspection of the results, the final step was to further validate our models on an entire cohort of students in the same way we intend to do in practice in the future. We did this with our fifth cohort of students, which was not used to train or cross-validate the models. Figure 2 shows these results in our 3-by-3 framework. AMMKNN(A) makes the best predictions: of 6 students who failed the PANCE, the model correctly predicted that 2 would fail, identified 2 as at risk, and failed to identify 2 (incorrectly predicting that they would score above 375). Of 5 students who passed with at-risk scores, the model failed to detect all 5.

Table 2. LOOCV Predictive Model Results Based on 2-by-2 Confusion Matrix

| Model (Variable Preparation) | TP  | FP  | TN  | FN  | Accuracy | Sensitivity | Specificity |
|-------------------------------|-----|-----|-----|-----|----------|-------------|-------------|
| AMMKNN (B)                   | 9   | 9   | 159 | 4   | 0.93     | 0.69        | 0.95        |
| AMMKNN (A)                   | 8   | 11  | 157 | 5   | 0.91     | 0.62        | 0.93        |
| SVM (B)*                     | 9   | 18  | 150 | 4   | 0.88     | 0.69        | 0.89        |
| Standard KNN (B)*            | 8   | 17  | 151 | 5   | 0.88     | 0.62        | 0.90        |
| RF (B)*                      | 10  | 23  | 145 | 3   | 0.86     | 0.77        | 0.86        |

Selected best models with each predictive approach are displayed; find all results in the Study Appendix (Supplementary Digital Content 1, http://links.lww.com/PAEA/A62). Example interpretation: LOOCV on AMMKNN with variable preparation A yielded 8 students as TP, 11 as FP, 157 as TN, and 5 as FN.

*These standard models required manual tuning of predicted results (after running the model) to be meaningful, which was not required for AMMKNN (see Study Appendix [Supplementary Digital Content 1, http://links.lww.com/PAEA/A62] for additional discussion).

AMMKNN, adaptive minimum match k-nearest neighbors; FN, false negatives; FP, false positives; KNN, k-nearest neighbors; LOOCV, leave-one-out cross validation; RF, random forest; SVM, support vector machine; TN, true negatives; TP, true positives.

Figure 1. LOOCV 3-by-3 confusion matrices for selected predictive models. Example interpretation: In the LOOCV results from adaptive minimum match KNN with variable preparation B, 11 students who were predicted to score above 375 actually scored between 350 and 375. AMMKNN, adaptive minimum match k-nearest neighbors; KNN, k-nearest neighbors; LOOCV, leave-one-out cross validation; SVM, support vector machine.
but this oversight would have been acceptable to us given that the students did not fail the examination. Finally, of 32 students who passed without risk, 2 were incorrectly predicted to fail, 0 were predicted as at risk, and 30 were correctly predicted to pass.

These results mean that if we had used this model in practice—ignoring other decision-making inputs in our discussion below—we would have remediated the 6 students classified as failing or at risk and we would have been right to remediate 4 of these 6 students who identified 4 of the 6 students who failed as either failing or at risk. Support vector machine and standard KNN failed to detect 4 and 3 students, respectively, of the 6 total who failed.

The results from both LOOCV and the entire-cohort validation suggest that AMMKNN might be more useful in practice than standard SVM and KNN models because (1) AMMKNN performs better in both A and B preparations of the independent variables and (2) AMMKNN has fewer false-positive predictions at desired levels of sensitivity.

**DISCUSSION**

Our results demonstrate (1) the potential value of an adaptive machine learning method such as AMMKNN that has high sensitivity while minimizing false positives, (2) the possible utility of categorizing students into different risk categories, and (3) that the predictions from machine learning models cannot be trusted blindly without considering other relevant factors. We argue that the AMMKNN model can only be useful in conjunction with other programmatic information to assist educators in identifying and supporting at-risk students. We argue that the use of predictive models in PA education needs to be considered from 3 perspectives—student, educator, and program administrator—which together can lead to a practical and ethically sound student support strategy. Balancing these perspectives involves trade-offs that remind us that analytics are best used as one among multiple inputs to decision making. We also acknowledge a number of strengths and limitations of our work in the Study Appendix (Supplementary Digital Content 1, http://links.lww.com/PAEA/A62).

**Student Considerations**

This study does not include an empirical examination of student reactions to receiving analytics-based results. We do, however, have previous experience in enrolling students in additional support for PANCE preparation, which gives us some of the insights that follow. Before incorporating predictive modeling, we identified at-risk students by using PACKRAT scores, a widely used approach. Naturally, many of these students were not enthusiastic to learn that they were considered to be at risk. Because the terms “at risk” or “remedial” might sound harsh, we might instead choose to use terms such as “needing extra support” or “supplemental curriculum.”

Nonacademic factors—such as personal, medical, and mental health issues that arise on or before examination day—are important to identify and acknowledge. We fear that students’ tendencies to “power through” the scheduled examination rather than rescheduling based on those circumstances could be detrimental in the long run. If a student wakes up on the examination day with a migraine and still decides to take the examination, our machine learning model can never account for this setback and adjust the prediction for that student. There will always be factors that are not included in the data. We recommend preparing students to be on guard for such circumstances so that they can reschedule their examination. Further research is required on student reactions to (their own) analytics results and nonacademic factors in PANCE performance.

**Educator Considerations**

Although educational analytics success stories abound, there is also concerning evidence that singling out students as being at risk can be problematic. Therefore, we argue that educators using this approach and similar predictive models...
need to be cautious. Although providing comprehensive recommendations for this process is beyond the scope of our current work, we note that established scholarship exists related to breaking bad news\textsuperscript{39,40} and remediation\textsuperscript{41,42} in contexts where students are preparing for high-stakes examinations. Framing\textsuperscript{43,44} the results when communicating with the students is something the educator must consider to ensure that students understand the motivation and uncertainties associated with suggested actions.

We recommend considering the model’s predictions along with other processes established by the program to ensure students’ success, including programmatic efforts to support students identified as having unmet needs for a variety of reasons. Typically, when we review the progress of our students each year and decide which students to recommend for our extra support program, we consider whether (1) a colleague has flagged a student for extra support, (2) a student has self-identified as needing more support, or (3) a student had a low PACKRAT score after their first year in the program. Now, we have an additional (fourth) tool: a machine learning model might identify (based on factors the human eye cannot see) additional students who did not have a low PACKRAT score, but who nevertheless fit the pattern of previous students who struggled. This tool gives us an opportunity to discuss those students in more depth than we would have otherwise. By combining information from all 4 of these approaches, we can then recommend extra support for a subset of students as they prepare for the PANCE.

Program Administrator Considerations

Once a process that incorporates analytics has identified at-risk students, program administrators still face the challenge of identifying and implementing successful interventions. A one-size-fits-all, extra-support intervention will not work for everybody, given the unique individual circumstances that students face (discussed above). How do we provide support that meets the needs of the students and stays within the realistic resources of a single program? Administrators must also be prepared for students who are not identified as being at risk raising concerns about equity. In other words, why should some students receive additional support for an incredibly high-stakes examination and not others? Administrators need to balance this equity issue against allocating additional attention selectively.

CONCLUSION

As educators, we hope to create a safety net around our students that can support them as needed, especially in PA education with patient outcomes at stake and delays in certification leading to unwanted professional consequences. The “adaptive minimum match” KNN model we have developed and validated serves as one additional tool that—along with already-existing tools—makes this safety net stronger when used responsibly and transparently. We were only able to test AMMKNN in a single PA program, and we recommend further research on this model and other methods across a variety of programs before the generalizability of AMMKNN can be known. We recommend that analytics should coexist with and bolster other approaches to student support, and that and that student, educator, and program administrator perspectives should all be incorporated for the practical and ethical implementation of educational analytics.

Anshul Kumar, PhD, is an assistant professor, MGH Institute of Health Professions, Department of Health Professions Education, School of Healthcare Leadership, Boston, Massachusetts.

Taylor DiJohnson, BA, is a project manager, Workplace Health and Wellness, Mass General Brigham, Human Resources, Occupational Health, Workers’ Compensation, Somerville, Massachusetts.

Roger A. Edwards, ScD, is a professor and chair, MGH Institute of Health Professions, Department of Health Professions Education, School of Healthcare Leadership, Boston, Massachusetts.

Lisa Walker, MPAS, PA-C, is a faculty member, University of Washington, MEDEX Northwest, Seattle, Washington.

Roger Edwards and Lisa Walker are both co-senior authors.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal’s Web site (www.jpae.pub).

We thank Rupali Khadye-Hadshi, Arimenta Russ, and Valay Maskey for their involvement in preparing and formatting this article for publication. We also thank the peer reviewers of this article not only for providing constructive critiques of the research that enabled us to strengthen this article but also for encouraging us to (1) think about how to make our writing more accessible and understandable to those who are new to machine learning and (2) identify ways to support institutions, programs, and fellow educators who are new to machine learning and are interested in initiating similar projects.

Correspondence should be addressed to: Anshul Kumar, PhD, Assistant Professor, MGH Institute of Health Professions, Department of Health Professions Education, School of Healthcare Leadership, 36 First Avenue, Charlestown Navy Yard, Boston, MA 02129. Telephone: (+1) 857.600.1864; Email: akumar@mghihp.edu

REFERENCES

1. Kucak D, Juricic V, Damigic G. Machine learning in education—a survey of current research trends. In: Katalinic B, ed. DAAAM Proceedings. Vol 1. 1st ed. DAAAM International; 2018:0406–0410.

2. Ekowo M, Palmer I. The Promise and Peril of Predictive Analytics in Higher Education: A Landscape Analysis. New America; 2016. https://files.eric.ed.gov/fulltext/ED570869.pdf. Accessed September 21, 2021.

3. Kumar A, Edwards R. AdaptiveLearalytics: Adaptive Predictive Learning Analytic Tools. GitHub; 2021. https://github.com/readcreate/AdaptiveLearalytics. Accessed May 15, 2023.

4. Kotsiantis SB. Use of machine learning techniques for educational proposes: a decision support system for forecasting students’ grades. Artif Intell Rev. 2012;37(4):331–344. doi: 10.1007/s10462-011-9234-x.

5. Stimpson AJ, Cummings ML. Assessing intervention timing in computer-based education using machine learning algorithms. IEEE Access. 2014;2:78-87. doi: 10.1109/ACCESS.2014.2303071

6. Akçaşpar G, Altun A, Aşkar P. Using learning analytics to develop early-warning system for at-risk students. Int J Educ Technol High Educ. 2019;16(1):40. doi: 10.1186/s41239-019-0172-z.

7. Sokkhey P, Okazaki T. Comparative study of prediction models on high school student performance in mathematics. In: 2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC). JeJu, South Korea; IEEE; 2019:1-4.

8. Liz-Domínguez M, Rodríguez MC, Nistal ML, Mikic-Fonte FA. Predictors and early warning systems in higher education: a systematic literature review. In: LASI Spain 2019. Vigo, Spain: Learning Analytics in Higher Education, 2019:84-99.

9. Aulck L, Velagapudi N, Blumenstock J, West J. Predicting student dropout in higher education. Paper presented at: 2016 ICML Workshop on #Data4Good: Machine Learning in Social Good Applications, June 24, 2016, New York, NY. doi:10.48550/arXiv.1606.06364.
