**Research problem (10pts): Describe the task you want to achieve. What is the outcome of interest? What are you trying to predict? Why is it important? What are the potential benefits of having a predictive model for this outcome? Discuss potential applications of such a model.**

Online learning has been a tool used for several years now; however, the coronavirus pandemic kick-started advancements in online technology and educational tools allowing students to continue their education from anywhere. This project explores student performance data taken from an online education management program. Student demographics as well as grade level, attendance rates, and other educational attributes are used to predict which category (low, medium, high final grade) students will fall into. The outcome of interest is which variable has the greatest effect on overall student performance. I predict overall successful performance will either be contingent upon the subject matter or student attendance. The subject matter has an effect on student mood and performance in school. Attendance is another factor that effects student’s interest in the subject as well as their overall performance.

The potential benefits for this type of model are to see where there are pitfalls in an online program. By using factors outside of student demographics, an online program would be able to observe what is effecting their student’s performance the most. Outliers or variables with significant negative effects can be brought forth to program developers for areas of improvement. This type of model could also be applied to in-person schools to observe any variables having a negative effect on overall student performance.

**Description of the data (15pts): Describe core features of the data, any additional features you produced from existing features and how, basic descriptive statistics about these features, and any missing data analysis you conduct. The description should be sufficiently clear that the instructor understands all the variables included in your modeling.**

The original dataset is housed on Kaggle (<https://www.kaggle.com/aljarah/xAPI-Edu-Data>) and was modified for the purposes of this project. This educational dataset was collected from Kalboard 360, a subscription-based, online learning management system that offers students access to synchronous classes from any device with an internet connection. This dataset was collected using a learner activity tracker tool called experience API (xAPI). “The xAPI is a component of the training and learning architecture (TLA) that enables to monitor learning progress and learner’s actions like reading an article or watching a training video. The experience API helps the learning activity providers to determine the learner, activity and objects that describe a learning experience” (Aljarah, I., 2016).

This dataset contains records from 480 students and 16 features as displayed in the table below. The original author of the dataset categorized these features into three main categories:

1. Demographic features (gender, nationality, etc.)
2. Academic background features (educational attainment, grade level, etc.)
3. Behavioral features (amount of times a student raises their hand or accesses resources)

Of the 480 students, 305 are male and 175 are female. Students’ nationalities breakdown into the following: 179 from Kuwait, 172 from Jordan, 28 from Palestine, 22 from Iraq, 17 from Lebanon, 12 from Tunis, 11 from Saudi Arabia, 9 from Egypt, 7 from Syria, 6 from USA, Iran and Libya, 4 from Morocco and one from Venezuela. Data was collected over the course of two semesters with 245 records from the first semester and 235 records from the second. It is interesting to note that a higher proportion of students missed less than 7 days of school during the semester. Finally, the dataset contains information gathered from parent surveys and parental level of school satisfaction. Overall, 270 parents answered the surveys and 292 parents noted their satisfaction with the school.

**Description of the models (15pts): List at least three different modeling approaches you apply to this dataset. Describe each model, why the given model was selected, which hyperparameters to be optimized and how. Also, discuss how you plan to evaluate model performance.**

I first chose a model using linear regression without any regularization; however, this was not chosen as it produced model fitting errors. Instead a model with 10-fold cross-validation using ridge regression was chosen to predict the class area students fell into. In order to find the optimal lambda value for evaluation, a k-fold cross-validation and plot were produced. Figure 1 displays the lambda score of 0.3673767 to be the best fit for all of the models. A model with 10-fold cross-validation using lasso regression was chosen as a comparison to the ridge regression model. Again, the optimal lambda score was used to produce the most effective prediction. Finally, a bagged tree model using a random forest was developed to predict the probability of student performance. Each of these models were chosen to determine the extent to which increasing the model complexity impacted the accuracy of the predictions. The mean-absolute error (MAE), root-mean squared error (RMSE), and the R-squared (Rsq) scores were calculated to compare the performance of each model.

**Model fit (20pts): Provide the results of your model evaluation. Compare and contrasts results from different fits, including a discussion of model performance. Discuss your final model selection and the evidence that led you to this selection. If it is a classification problem, how did you choose a cut-off point for binary predictions? Did you consider different cut-off points?**

Table 2 provides a cross-comparison of models using their MAE, RMSE, and Rsq scores. The RMSE score for the bagged tree model using a random forest was the lowest out of the three models. Its RMSE score of 2.044026 describes the standard deviation of the differences between the predicted and observed values. The random forest model produced an R-squared score of 0.6727365, so about 67% of the model’s predictions are correct. This R-squared score was the highest out of the three models, which displays it is the best at predicting student performance. This model could be improved to produce a higher R-squared score by scaling features and continuing to tune hyperparameters.

**Data visualization (5pts): Include at least two plots (or more) to help communicate your findings. The plots may be of initial data explorations, fits of individual models, and plots displaying the performance of competing models.**

During initial data exploration, a histogram was created to check whether the dependent variable (Class) followed a normal distribution pattern. The highest amount of students observed falls in the middle rather than on either side indicating a normal distribution pattern.

A correlation matrix was also developed to determine which variable was the most important in predicting the outcome. A student’s performance and grade (Class) was most highly correlated with the amount of times they raised their hand during instruction.

**Discussion/Conclusion (25pts): Discuss and summarize what you learned. Which variables were the most important in predicting your outcome? Was this expected or surprising? Were different models close in performance, or were there significant gaps in performance from different modeling approaches? Are there practical/applied findings that could help the field of your interest based on your work? If yes, what are they?**

Within this project, I tested five models before choosing the best three. Both the model using linear regression without any regularization and bagged tree model did not produce favorable outcomes. All three models performed to similar standards with no significant gaps in performance. Each model produced MAE, RMSE, and R-squared scores that were decimals away from one another. From the three models observed, they all predicted the outcome with over 65% accuracy. This percent would need to be improved if the model were to be used to collect data used to improve online schools. As noted above, this could come from scaling features and continuing to tune hyperparameters or it may be a matter of choosing a different model type.

The amount of times a student raised their hand during instruction had the highest correlation to student performance overall. This was surprising to me as my hypothesis was that the class subject or student attendance would have the highest correlation. However, it shows that this online program does an excellent job of engaging their students in the material. An engaging classroom would hold students’ attention and create peer collaboration fostering deeper learning which, in turn, would lead to higher student success. This online program should be praised for its commitment to student participation. While the model’s prediction accuracy does need to be improved, there is promise for practical application in the future. A model that uses student learning characteristics to predict their overall performance could be used several times through the school year to continuously be evaluating areas of improvement.

Due to the pandemic, online classes and schools have become a more commonly utilized tool. Individuals are able to receive an education at any stage in life and anywhere geographically around the globe on their time schedule. More individuals are prone to obtain a degree online due to the drastic advancements in educational technology. If this trend is to continue, an evaluation metric will need to be developed in order to accommodate online students. A prediction model, like the one developed for this project, is a sustainable and simple way to evaluate online programs based on predictions of student success.

**Reproducibility (10pts): Provide a link to the GitHub repo at the beginning of your report as a note.**

**References**

Aljarah, I. (2016). Students' Academic Performance Dataset. In *Kaggle*. https://www.kaggle.com/aljarah/xAPI-Edu-Data

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining Educational Data to Predict Student’s academic Performance using Ensemble Methods. International Journal of Database Theory and Application, 9(8), 119-136.

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2015, November). Preprocessing and analyzing educational data set using X-API for improving student's performance. In Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on (pp. 1-5). IEEE.