Research Question:

- *Is it possible to predict a child's class level based off of the information provided in the dataset, and if so, which variables were of significance?*
- **Dataset:**
    - 480 Observations
    - 17 Variables: *Gender, Nationality, Place of Birth, Stage ID, Grade ID, Section ID, Topic, Semester, Relation, Raised Hands, Visited Resources, Announcements View, Discussion, Parent Answering Survey, Parent School Satisfaction, Student Absence Days,* ***Class***
- **Explanatory Variables: Any combination of the variables listed above.**
- **Response Variable: Class (3 Levels: High, Medium, Low)**

Table of Coefficients:

Term	Estimate	P-Value
Gender (Male)	-4.379	.009
Semester (Spring)	1.752	.15
Visited Resources	0.062	.002
Announcements View	0.060	.027
Parents Answering Survey (Yes)	2.231	.044
Student Absence Days (Under 7)	5.780	.00006

This is a table of coefficients gathered from our Backwards Stepwise Logistic Regression model, which was deemed to be the most appropriate for answering our research question.

Conclusion

- Ultimately, we were able to prove that it is possible to predict a child's class level based off of the information in the dataset, and found several variables that were sufficient in prediction.
- These included: Gender, Semester, Visited Resources, Parent Answering Survey, Viewed Announcements and Student Absence Days.
- Utilized the following statistical methods in our analysis:
    - KNN model utilizing all 3 levels of the Class variable (60% prediction accuracy)
    - KNN model using only 2 levels of the Class variable with the same predictor variables (low and high) (94% prediction accuracy)
    - Backward stepwise selection model, originating with all but 3 variables (96% prediction accuracy).
    - McFadden’s pseudo-R-squared value in the stepwise logistic model was the highest out of our logistic models (0.9) respectively thus showcasing that the stepwise model was, infact, the strongest model.
- This study could be improved by using a multinomial regression model, as well as KNN using Manhattan distance rather than Euclidean. Finally, using a different goodness of fit test would greatly aid the validity of our results.