# ## STAT 4185 Final Project Analysis

Process:

- Data Collection: load dataset

- Data Preprocessing: cleaning and splitting with Pandas, scaling

- Data Visualizations with matplotlib and seaborn (will come back to this)

- Model Construction: decision tree/random forest

- Model Assessment: predictive performance analysis

## Data Collection

I went about seeking datasets on Kaggle because it serves as a hub of repositories containing various Jupyter notebooks and projects, and stumbled upon a study from a Portuguese educational institution. This dataset consists of student information such as the course (major, area of study),

Works Cited

GeeksforGeeks. “Create a Stacked Bar Plot in Matplotlib.” *GeeksforGeeks*, 24 Aug. 2022, www.geeksforgeeks.org/create-a-stacked-bar-plot-in-matplotlib.

“Matplotlib Different Size Subplots.” *Stack Overflow*, stackoverflow.com/questions/10388462/matplotlib-different-size-subplots.

“One-Hot Encoding of Label Not Needed?” *Stack Overflow*, stackoverflow.com/questions/62911538/one-hot-encoding-of-label-not-needed.

“Predict Students’ Dropout and Academic Success.” *Kaggle*, 3 Jan. 2023, www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention/data.

Realinho, Valentim, et al. “Predicting Student Dropout and Academic Success.” *Data*, vol. 7, no. 11, Oct. 2022, p. 146. https://doi.org/10.3390/data7110146.

“Rotate Axis Tick Labels.” *Stack Overflow*, stackoverflow.com/questions/10998621/rotate-axis-tick-labels.