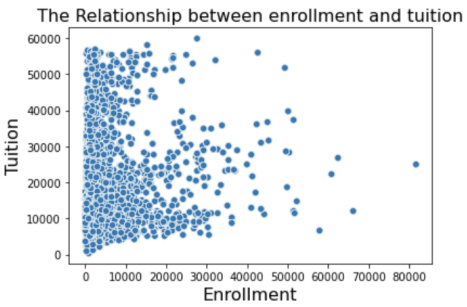
FP3 – Data Analysis Plan

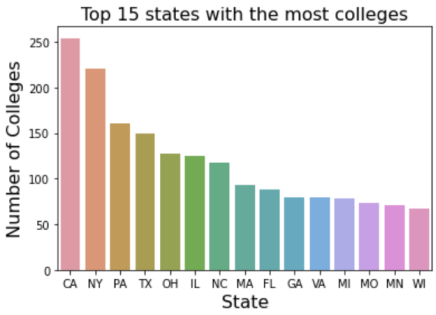
Overview

- ✔ The topic of our project relates to colleges in America. Nowadays it is almost essential to go to college if one wants to have a step up in life. A college degree is seen in our society almost as a commodity nowadays, and carries much more social value than just knowledge and education. For our project, we will try to classify colleges based on criteria that a prospective student might be looking for. For a student who does not know which college they want to go to, but has an idea of the enrollment size they would like, the ideal tuition cost, or type of school they want to attend, our project could be of benefit to them. They will be able to input these criteria and our program will predict the best institution options for them.
- ✔ We believe it's important to tackle this problem because it is relevant to us as students along with the whole country as a system. Some people in our world don't have access to as many resources to be as knowledgeable about the college process as others. Ideally, our program would be a resource for anyone with internet access to learn more about the best colleges for them.
- ✔ Our dataset includes college name, state, type, degree length, in state tuition, in state total, out of state tuition, out of state total, total enrollment, category, and enrollment by category. There is a mix of categorical variables and cardinal variables. Our data set has thousands of rows so we feel that we have enough data.
- ✔ Our data set includes the name of the college as a target variable and the rest of the variables as feature variables. Given some of the feature variables such as out of state tuition, state, and total enrollment, we could have an algorithm try to predict the name of the school. These features will be useful in predicting our target variable because these are each important aspects in picking a school: tuition, location, total enrollment, etc.
- ✔ How do the k-nearest neighbors, support vector machine, and decision tree algorithms compare in correctly predicting a college based on the given features?
- ✔ How accurate can we be when classifying a college? Are there too many college names for the algorithm to correctly classify each one?

Data Analysis Plan

- ✔ Our project is tackling a classification problem. Given variables such as State, degree length, in state tuition, out of state tuition, and total enrollment, we want to be able to classify the college. A hypothetical use of this could be where a high school student does not know where to apply to college, so they can input data and our analysis can give a college based on their data.
- ✔ The algorithms we will use will be k-nearest neighbors, support vector machine, and decision tree. We know that the k-nearest neighbors algorithm is easy to understand and the model is built pretty fast. However, we know that it may be slow for large training datasets, which ours is.
- ✔ We know that a support vector machine algorithm has fast prediction and scales well with size. However we know that other classifiers might be better for low dimensional datasets.
- ✔ We know that decision trees are easily visualized and work well with datasets that contain a mixture of feature variables which our dataset does. However we know that decision trees sometimes still fail to generalize and may not perform as well as other algorithms.
- ✔ At first glance, we personally think that decision trees would have the highest performance, but we still don't really know. Because of this, we will be certain to try each of the algorithms to see how our model performs. In addition, we will try to do hyper parameter tuning to increase performance.
- ✔ For feature engineering, we will try to use one-hot encoding because it is a good way to represent categorical data, which we have a lot of.
- ✔ We thought it would be useful to visualize how many colleges are there in each state by using a map where a darker shade means there are more colleges there. We also wanted to look at how tuition compares with enrollment. We could achieve this by doing a scatter plot with enrollment on the x axis and tuition on the y axis.





- ✔ The visualization on the left is a scatterplot which compares enrollment to tuition. Although there isn't necessarily a strong linear relationship or pattern, we can somewhat see that for a few schools, tuition price does increase very slightly as enrollment gets bigger. However, for most schools it is very scattered and tuition doesn't really depend on enrollment size.
- ✔ The visualization on the right is a barchart where the states are on the x axis and the y axis has the number of colleges in that state. This matters because there are more schools in California, New York, and Pennsylvania then compared to Vermont for example. This means that assuming our inputs were given at random, then a school from California is more likely to appear because there simply are just more schools in California.