**SVM**

**Data Cleaning**

The timesData dataset required the following data cleaning procedures:

1. Filtering the dataset for 2016 year.
2. Dropping the world\_rank column for fitting support vector regression and dropping total\_score for fitting SVC and SVM.
3. Removing rows with missing and NaN values.
4. Removing rows with hyphen values from income feature column.
5. Removing the commas in integer values in thousands separator in num\_students column and hence, converting from string to float.
6. Replacing the hyphens in total\_score column with a float value of 45 since all the values in this column below 48.8 were missing and denoted by hyphens.
7. Removing the % sign from international\_students column and hence, converting from string to float.
8. Removing rows with hyphen values in female\_male\_ratio column.
9. Converting the female\_male\_ratio column from ratio (string) to a decimal value (float).
10. Dropping the university\_name from model prediction because all the rows have unique feature values and no two rows have the same values for which the university\_name column feature could separate them from one another.
11. Creating dummy variables for country column for analysis.
12. Scaling the features using MinMax scaler.

1.b) When we fit support vector regressor, we observe the test scores are negative which implies that SVR is not really helping us predict data on unseen datasets accurately.

1.d) When we perform cross validation with various values of C for SVC model, we observe that accuracy of the model increases as we increase C, this is consistent with the fact that as we increase C, we increase the penalty and allow for a smaller margin model which decreases the misclassification. This follows the fact that the number of training errors’ misclassification decreases with an increase in C. The cross validation error rate also decreases with an increase in C as when the number of training errors misclassified decreases, the cross validation accuracy increases and eventually leads to a decrease in cross validation error rate.