ML Model Checklist

Frame the problem

Can we predict students' performance in academics in STEM and Non-STEM?

Is there a difference in STEM and NON-STEM when looking at the performance of students and their features?

Select a performance measure

Ridge and Lasso Regression:

MSE for less outliers

MAE for outliers present

Set up a virtual environment

Set up libraries on virtual env; matplot, seaborn, pandas, numpy, scikitlearn

Exploratory Analysis and Visualization

- Histograms
- Describe function
- Null values
- Outliers
- Data types
- Seaborn heatmap of co correlations

Make training and test sets

20/80 split - scikit learn built in function

Data Wrangling

- Handle missing values; replace them with mean/median, get rid of them
- Convert categorical variables using one hot encoding
- Feature Scaling - scaling target label not required. Other labels should be scaled as ML algorithms do not learn well with large variances in scales of features. Standard scaler of MinMaxScaler

Select and Train Model

Perform k fold cross validation

Ridge Regression
Lasso Regression
Decision Tree Regressor
Random forests Regression

Fine Tune Model

Use grid search if sample space is smaller otherwise randomized search.

Check feature importance of each attribute and remove features that are less impactful on the models.

Use Model on Test Set

Transform data on the test set for the model
Avoid tuning hyperparameters to better fit the test set - THIS DOES NOT GENERALIZE WELL

Analyze and Present Results