# Young Final Project Proposal

# Table of Contents



---

## Choose dataset

NASA Hazardous Asteroid Classification dataset from Kaggle (https://www.kaggle.com/shrutimehta/nasa-asteroids-classification)

## Big question

How few features can be used to accurately predict a hazardous asteroid versus a safe asteroid?

## Process

| To Do | Description | Goal | How To |
|---|---|---|---|
| Preprocess | Load data, clean data, encode if necessary | Have a clean dataset | Load data properly with Pandas; use `df.info()`, `df.describe()`, `sum(df.isna().any())`, etc to understand data and find what needs to be cleaned; remove nulls, replace 'dirty' data, change `dtypes`; use `pd.get_dummies()` to encode target values to 0/1 if necessary |
| Visualize | Plot features in data to derive basic insights of features | Find important/effective features to use in model as well as features to possibly exclude | Use different plots in seaborn, matplotlib, plotly, etc to derive insights from different features. `seaborn.pairplot()` might be a good place to start |
| Split data and search for best model | Define X (variables) and y (target), split into Train/Test or Train/Dev/Test set, choose model(s) to use, set up parameters, run parameters and train set through GridSearchCV | Have a strong model to build defined by our limited search | RandomForestClassifier, KNN, etc and GridSearchCV |
| Build and score model | Use parameters defined by highest scoring model found in GridSearchCV to build model | Successfully send data into model and get a score in some form | Build classification object with parameters, fit on train data, plot confusion matrix and get classification scores |
| Train other models | There might have been other models not used. Spend some more time training other models that work (see score in GridSearchCV or choose another model) | See if another model works better | Repeat building and scoring with different models, teasing hyperparameters to see if we can get a better result |

# Presentation

8 - 10 pg report. I know it's not necessary, as I am in the 475 class, but I want to work on my report writing skills, especially as a potential graduate program opportunity is in the future.

# 3 Skills to Demonstrate

1. Automation
2. Informative plots
3. Can I utilize new (not learned yet in class) packages to achieve a better score?