# NASA Hazardous Asteroids Classification

Alison J. March
Data Science Department
University of Colorado, Boulder
Boulder, CO 80309-0552
alma2157@colorado.edu

## Abstract
The goal of this project is to study and explore effective data mining and Machine Learning methods to identify and classify hazardous asteroids. In addition, we will compare those methods and determine the best machine-learning models for this project.

## 1. Introduction
The dataset is provided by NASA's Near-Earth Object Observations Program via NeoWs (Near Earth Web Service) Restful API web service for near-earth asteroid information(https://api.nasa.gov/). All the data is maintained by the NASA JPL Asteroid team(http://neo.jpl.nasa.gov/). It consists of 40 data features(columns) and 4,686 data instances (rows) of asteroid entries [1]. We reviewed and pre-processed the dataset to eliminate unnecessary features before using Machine Learning models to complete the classification task.

## 2. Related Work
So far, we have exported the dataset and done an initial data review and exploratory data analysis with Jupyter Notebook. Redundant features like ***Est Dia in M(min)*** and ***Est Dia in M(max)***, ***Est Dia in Miles(min)***, ***Est Dia in Miles(max)***, ***Est Dia in Feet(min)***, and ***Est Dia in Feet(max)*** were dropped. Furthermore, features like ***Close Approach Date***, ***Orbiting Determination Date***, ***Orbiting Body***, and ***Equinox*** were eliminated before Feature Engineering and Feature Selection. A correlation heatmap is also constructed to study the correlation between data features.

## 3. Proposed Work
### 3.1 Dimensionality Reduction
For the remaining 30 data features, we will use the Chi-squared based SelectKBest library to rank each feature based on the correlation score: a high score represents a high correlation and vice versa. We want to drop features with a correlation > 90% to prevent multicollinearity. Min-max scaling normalization will also be carried out to transform the dataset. Finally, we will check and resolve any data imbalance in preparation for Machine Learning algorithms.

### 3.2 Supervised Machine Learning Models
#### *3.2.1 Logistic Regression*
#### *3.2.2 Decision Trees*
This section will be updated once we have trained and evaluated all the models and have stats and insights.

#### *3.2.3 Support Vector Machines*
This section will be updated once we have trained and evaluated all the models and have stats and insights.

#### *3.2.4 Random Forest*
This section will be updated once we have trained and evaluated all the models and have stats and insights.

#### *3.2.5 XGBoosting*
This section will be updated once we have trained and evaluated all the models and have stats and insights.

## 4. Evaluation
We will split the cleaned dataset in the ratio of 80/20 train/test sets. Cross-Validation is a good way to evaluate a model by using Scikit-Learn's `cross_val_score()` function. Tentatively we will choose the StratifiedKFold class to perform stratified sampling to produce folds that contain a class ratio. The process behind such a method is to create a clone of the classifier at each iteration, train a clone on the training folds and output a prediction on the test fold, and count the number of correct predictions and the ratio of correct predictions. A `precision_recall_curve()` function can be utilized to visualize the precision and recall versus the decision threshold. The second method to evaluate the model is to compute the confusion matrix, this can be done with Sciki-Learn's `cross_val_predict ()` function. It provides precision and recall scores, and we can combine these two into a single metric called $F_1$ score.

## 5. Discussion
This section will be updated once we have trained and evaluated all the models and have stats and insights.

## 6. Conclusion
This section will be updated once all analyses and results have been collected and documented.

## 7. References
[1] Shruti Mehta. 2018. NASA: Asteroid classification. (March 2018). Retrieved May 20, 2023, from https://www.kaggle.com/datasets/shrutimehta/nasa-asteroids-classification