# PaAC Asteroid Diameter Prediction Progress Report

**Name -** Anikesh Parashar
**Enrollment No. -** 21114012
**Email ID -** a_parashar@cs.iitr.ac.in

## Project Introduction

- The Open Asteroid Dataset which is maintained by the Jet Propulsion Laboratory (JPL) of NASA has been used for the project.
- The Random Forest Regression, Extra Trees Regression, AdaBoostRegressor, Gradient Boosting Regression, HistGradientBoost Regression, Support Vector Regression, XGB Regression, LightGBM Regression and Multi-Layer Perceptron Regression were used for model fitting.
- The performance metrics used for comparing the models on the basis of test results were:
    - Mean Absolute Error
    - Mean Squared Error
    - Median Absolute Error
    - Explained Variance Score
    - r2-Score
- Apart from this, the models were also compared on the basis of Cross Validation using the following metrics:
    - Scores
    - Mean
    - Standard Deviation

## Weekly Progress

### Week 1

- Imported required libraries
- Read ‘Asteroid_Updated.csv’ file
- Viewed general characteristics of the data, like number of records and data types of individual columns or features
- Converted diameter from string to float
- Dropped unrequired columns like name, extent, rot_per, GM, BV, UB, IR, spec_B, spec_T and G
- Found correlation between different features and plotted the heatmap.
- Used Kernel Distribution Estimation (KDE) to obtain the distribution of various data

- Plotted the graph between various features vs diameter, log of features vs diameter and features vs log of diameter
- Used one hot encoding on condition_code, neo, pha and class
- Used KNN imputation using all non-null features to fill null values in data_arc, H and albedo
- Realised that H and log of diameter had a linear relation
- Determined approximate relation between diameter and H
- Introduced a new feature exp_neg_kH, derived from feature H
- Plotted a new heatmap including the new feature.

## Week 2

- Found correlation of diameter with nth power of each numerical feature. (n varies from 1 to 10, and -9 to -1)
- Realised it would be better to convert condition_code to float and use ordinal encoding on class
- Realised it would be better to use KNN with features that have high correlation with features having null values, rather than using all features
- Tried to find relation between other data
- Introduced new feature b
- Found correlation of diameter with (1/n)th power of each numerical feature. (n varies from 2 to 5, and -2 to -5)
- Found correlation of diameter with product and quotient of each pair of features
- Dropped column n_obs_used

## Week 3

- Started to analyse relation between diameter and various features
- Calculated Spearman and Kendall correlation coefficients
- Dropped column per_y because of its high correlation with per
- Manipulated features which show low correlation with diameter so that their distribution becomes Gaussian
- Dropped less important features showing high correlation with other features. These include a, q, ad, per, e, pha and data_arc
- Added new features ekH_div_albedo, ekH_neo and ekH_pha because they have maximum correlation coefficients
- Standardised data using Standard Scaler
- Saved DataFrame as ‘Asteroid_Modified.csv’. The data in this file is used to train models
- Read data from ‘Asteroid_Modified.csv’ and splitted Test and Train Data
- Started fitting basic ML Model Ensembles on the features present in the data set

## Week 4

- Read data from ‘Asteroid_Modified.csv’ and splitted Test and Train Data
- Trained a few basic DNNs

- Played around with the basic features like number of neurons, learning rate, decay rate etc.
- Compared the scores of various networks and found the best one
- Used RandomizedSearchCV to find a good Neural Network

# Result

The Random Forest Regressor gave the best results, with the following metrics.

## Cross Validation

- Scores: 0.68802357, 0.28007676, 0.06561844, 0.0558667, 0.04707463, 0.03946168, 0.03526181, 0.04808714, 0.06202095, 0.06453636
- Mean: 0.13860280285633117
- Standard deviation: 0.19557082408696733

## Test Data

- Mean Absolute Error : 0.04189141065507549
- Mean Squared Error : 0.06317341583802455
- Median Absolute Error : 0.02113960507459816
- Explained Variance Score : 0.9432657426983421
- r2-Score : 0.9432657325718565