

# MODEL REPORT

Team Name: Code Blue

## MODEL USED

---

- Bagging Classifier
- On Decision Tree

## SCORE

---

- Cross\_val\_score:  
Avg of 99.991%
- Model Score:  
99.35 %
- Parameters Set:  
n\_estimators = 100  
max\_samples = 85%

## RELATIONS USED

---

- $a = q/(1-e)$

## TRAINING DATA ANALYSIS

---

### Removed Columns:

- ~ Name
- ~ Diameter

(As there are many datapoints in the training dataset I considered it's just better to remove the parameter than interpolating it, also I thought that its irrelevant with prediction we are doing as diameter seems to contribute solely to the size of the asteroid)

- ~ Albedo (Same reason as diameter)

- ~ Rotation Period

### Handling NaN and anomalous Values:

- ~ "a"

There's one row with  $e = 1$ , meaning it's a parabola hence i filled 'a' in that row with a very large value of  $10^{10}$  (1e10) since the model can't take infinity as input in dataset.

As there were some 'e' values greater than 1 there were some negative 'a' values implying those have hyperbolic orbits

I considered them to be valid datapoints as they tell the model about hyperbolic orbits even though negative of 'a' is not possible.

- ~ "neo"

Can be estimated by albedo but most of the data in albedo is Nan, So I removed the rows. (Only some are missing so I think it doesn't make a huge difference)

- ~ "h"

Can be calculated by albedo and diameter but mostly data in diameter and albedo is Nan, So I removed the rows. (Only some are missing so I think it doesn't make a huge difference)

- ~ "Coditional\_code & Data\_arc"

I think it can't be calculated using any data, so removed the rows. (Only some are missing so I think it doesn't make a huge difference)

### Replaced 'n' as 0 and 'Y' as 1

---

## TEST DATASET

---

### Removed Columns:

- ~ Name
- ~ Diameter

### Handling NaN and anomalous Values:

- ~ "a"  
Used the relation of  $q/(1-e)$  to fill the Nan
- ~ "ad"  
Used the relation  $a(1+e)$  to fill the Nan
- ~ "Data\_arc & H"  
Filled by taking the mean of the data
- ~ "Conditional Code"  
Left with Nan values as our model could take nan in conditional code Column

## MODEL TRAINING

---

Used cross\_val\_score to check models best average score and for checking if the model is generalizable and is not getting overfitted.

Model Checked: SVC, Decision Tree, Logistic Regression, Bagging on Decision Tree, Random Forest and Gaussian Naïve Bayes.

Split the training data into 0.75 train and 0.25 test dataset for checking the models working and to not overfit it with the data.

Trained Random Forest and Bagging in Decision Tree with the split train dataset (As they have less variation between cv splits in cross\_val and have high average scores.

**Selected Bagging on Decision Tree as it had the maximum score.**

---