**Introduction**

In this analysis, we evaluate the performance of three machine learning algorithms, namely Logistic Regression, Random Forest, and Decision Tree, on our dataset. The evaluation will include metrics such as R^2, MAE, MSE, and RMSE to provide insights into the predictive capabilities of each model. We also plot shapely plot and analyse data.

**Dataset**

The dataset is retrieved from Github [Predicting-Asteroid-Diameter-Dash/Pred\_Ast\_Diam\_2.csv at master · blakelobato/Predicting-Asteroid-Diameter-Dash · GitHub](https://github.com/blakelobato/Predicting-Asteroid-Diameter-Dash/blob/master/model/Pred_Ast_Diam_2.csv) (Blakelobato, 2020b). It consists of 23 features and 126497 rows of data. The 23 features are:

**diameter:** This is the target variable for this project. It is in km for the unit.

**orbit\_id:** Orbital solution ID

**e:** Eccentricity of the orbit

**a:** Semi-Major Axis in units (AU)

**i:** Inclination; angle with respect to x-y ecliptic plane (deg)

**om:** Longitude of the ascending Node (deg)

**w:** Argument of Perihelion (deg)

**ma:** Mean anomly (deg)

**n:** Mean Motion (deg/d)

**tp:** Time of perihelion passage

**moid:** Earth minimum orbit intersection distance (au)

**moid\_jup:** Jupiter minimum orbit intersection distance (au)

**class:** Orbit Classification

**producer:** Name of Person (or Institution) Who Computed the Orbit

**data\_arc:** Number of days spanned by the data-arc (days)

**n\_obs\_used:** Number of observations used in the orbit fit (all types)

**rms:** Normalized Root Mean Square of orbit fit (arcsec)

**albedo:** Geometric albedo

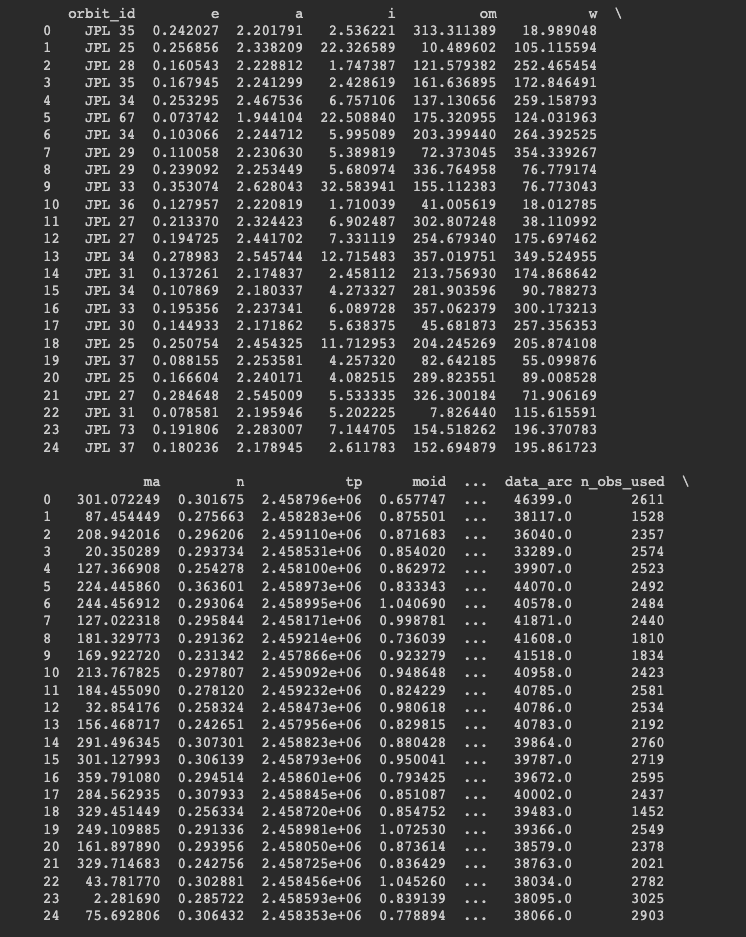
**first\_year\_obs:** The year of the first observation

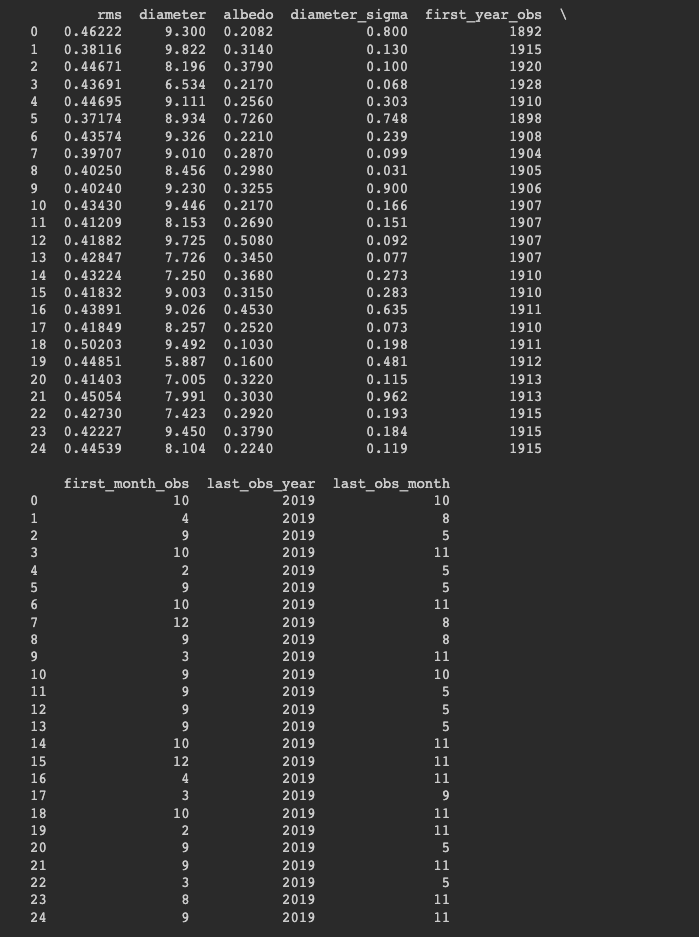
**first\_month\_obs:** The month of the first observation

**last\_year\_obs:** The year of the last observation

**last\_month\_obs:** The month of the last observation

Data form the top 25 rows are:





**Data Cleaning**

For this assessment the data was already cleaned all outliers were removed, all the missing values were eliminated, and all the irrelevant data were removed.

Visualising the distribution of the target variable

A picture containing screenshot, display, text, software

Description automatically generated

Figure 1 Distribution of target variable (Diameter) before data cleaning

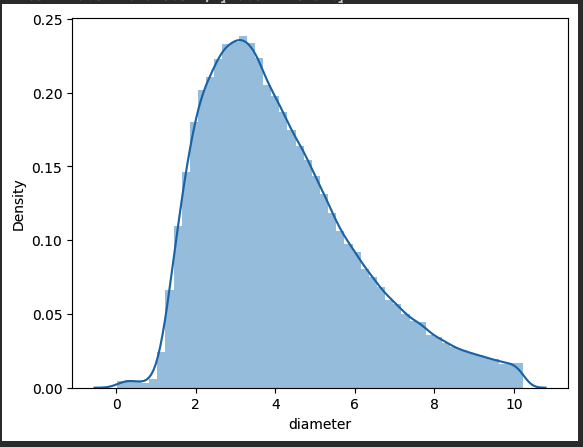


Figure 2 Distribution of target variable (Diameter) after data cleaning

This plot provides us with important information about the distribution of asteroid sizes. Here's what we can understand from it:

Most Common Sizes: The highest point on the curve represents the most common asteroid sizes. In this case, it occurs when the x-value is around 3.8, and the y-value is around 0.23. This suggests that there are more asteroids with sizes around this value compared to other sizes.

Spread of Sizes: The curve shows us how the asteroid sizes are spread out. If the curve is broader, it means the sizes are more spread out, and if it is narrower, it means the sizes are more concentrated around the most common size.

Rare Sizes: As the curve drops down towards 0 on the y-axis, it indicates that larger asteroid sizes become increasingly rare. This suggests that there are fewer asteroids with very large sizes.

The smoothness of the line connecting the points on the plot indicates that the distribution of asteroid sizes is relatively continuous and doesn't have sudden jumps or gaps between different sizes.

**Correlation Matrix**

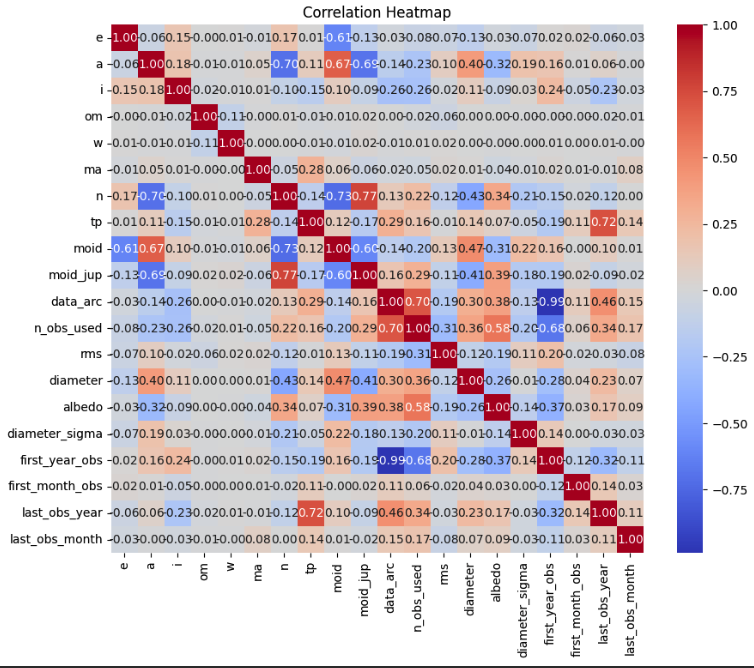


Figure 3 correlation Heatmap of all the features

Based on the correlation coefficients the attributes that show relatively stronger correlations with the diameter of the asteroid are:

moid (Earth minimum orbit intersection distance): This attribute has the highest positive correlation coefficient (0.47) among all the attributes. It represents the minimum distance between the asteroid's orbit and the Earth's orbit. A larger value of moid indicates a closer approach to Earth, and it correlates with a larger diameter of the asteroid. This attribute seems to have a significant impact on the asteroid's size.

a (Semi-Major Axis): This attribute has a relatively strong positive correlation coefficient (0.40) with the diameter. The semi-major axis represents the average distance of the asteroid from the Sun. As the semi-major axis increases, indicating a farther distance from the Sun, the diameter of the asteroid tends to increase as well.

n\_obs\_used (Number of observations used in the orbit fit): This attribute shows a positive correlation coefficient of 0.36 with the diameter. It represents the number of observations used to determine the asteroid's orbit. A higher number of observations used in the orbit fit is associated with a larger diameter of the asteroid.

data\_arc (Number of days spanned by the data-arc): This attribute has a moderate positive correlation coefficient of 0.30 with the diameter. It represents the duration of time covered by the observations used to determine the asteroid's orbit. A longer data arc, indicating a greater time span of observations, tends to be associated with a larger diameter of the asteroid.

**Scatter Plots**

We take in three important features from the data set and plot a scatter plot against the target variable.

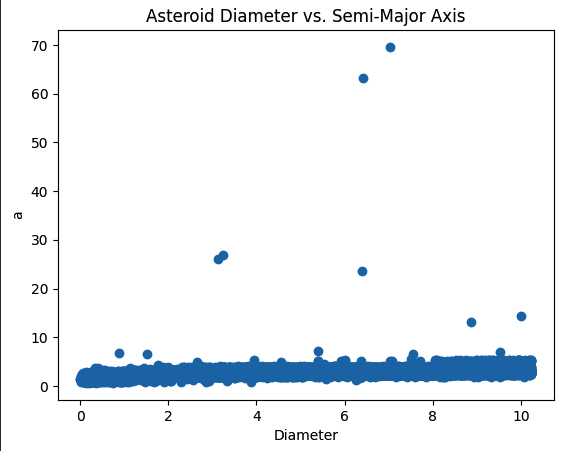


Figure 4 Scatter plot Asteroid Diameter VS Semi-Major Axis

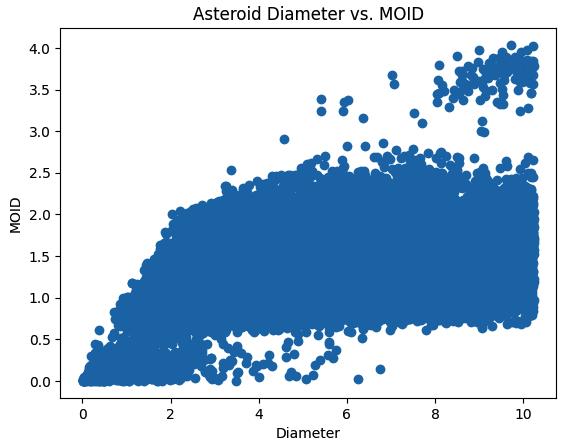


Figure 5 Scatter Plot Asteroid Diameter VS Moid

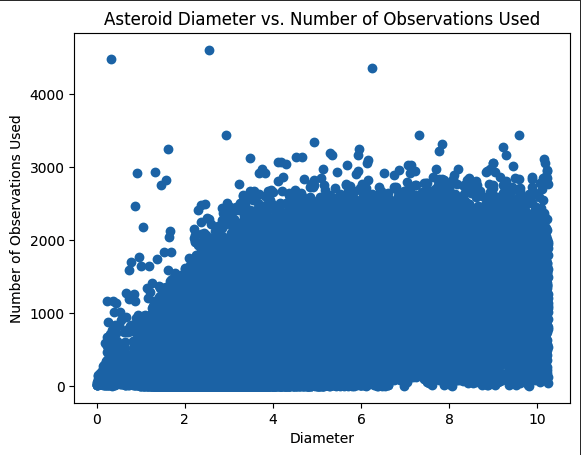


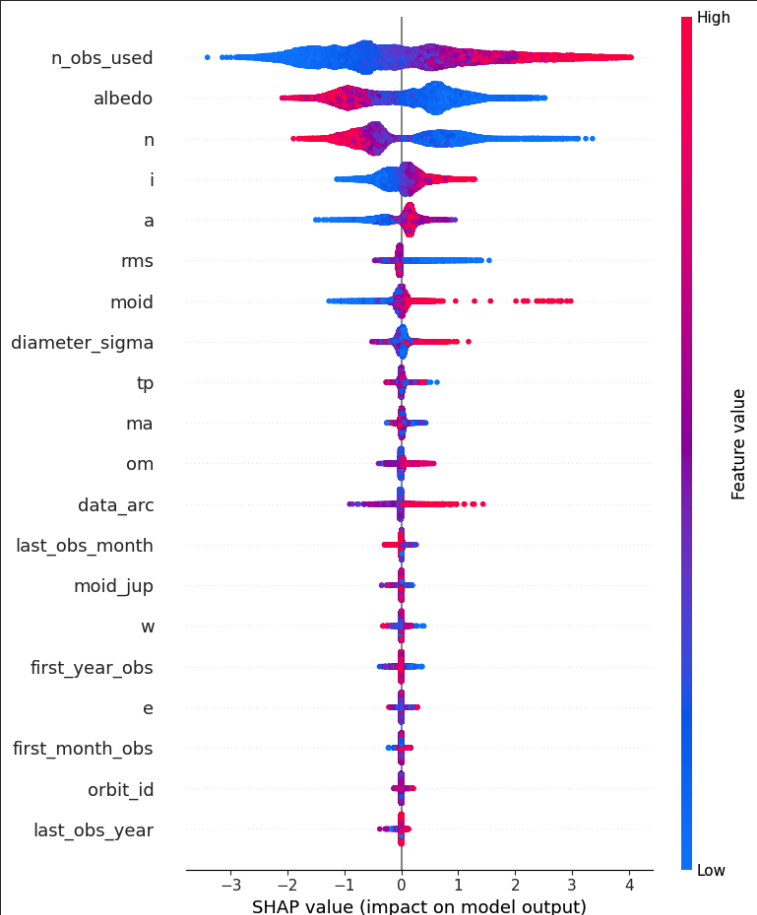
Figure 6 Scatter Plot Asteroid Diameter VS Number of Observations

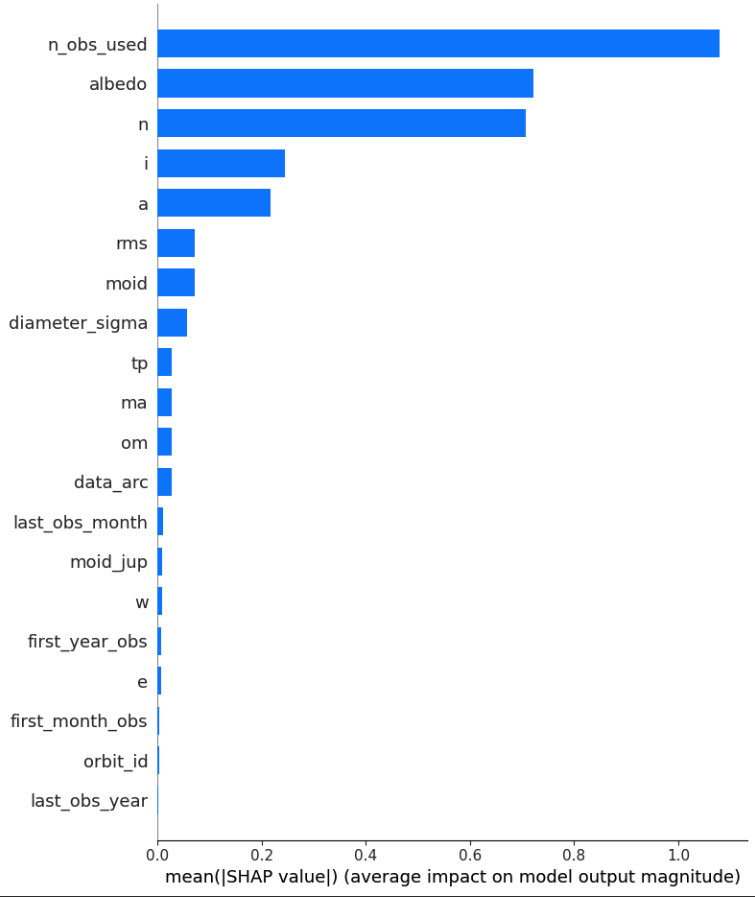
**Model Implementation**

|  |  |  |
| --- | --- | --- |
| **Models** | **Performance or Evaluation Metrics Root-Mean-Square Error (RMSE)** | **Evaluation Comment** |
| Logistic Regression | * Linear Regression, dependent on ['n\_obs\_used']: * Training MAE Error: 1.45 km standarized * Training MSE Error: 3.26 km standarized * Training R^2 Error: 0.1281 * Validation MAE Error: 1.45 km standarized * Validation MSE Error: 3.22 km standarized * Validation RMSE: 1.93 km standardized * Validation R^2 Error: 0.1363 | * The model performs consistently on both the training and validation sets, as indicated by similar MAE and MSE errors. * The R^2 error is low, suggesting that the model explains only a small portion of the variance in the target variable. * The RMSE error on the validation set is 1.93 km, indicating the average prediction error of the model. |
| Decision Tree | * Training R^2 value 0.9193811733358744 * Validation R^2 value 0.854488266055443 * Training MAE: 0.3909096188413319 km (standarized) * Validation MAE: 0.5274255726226369 km (standarized) * Training MSE: 0.30161600917217257 km (standarized) * Validation MSE: 0.5431147314904761 km (standarized) * Training RMSE: 0.5491957840080098 km (standardized) * Validation RMSE: 0.7369631819097044 km (standardized) | * The model shows high R^2 values for both the training and validation sets, indicating a good fit to the data. * The MAE and MSE errors are relatively low, suggesting that the model has a small average prediction error. * The RMSE error on the validation set is 0.74 km, indicating the average prediction error of the model. |
| Random Forest | * Training R^2 value 0.955232048436248 * Validation R^2 value 0.9058072084354329 * Training MAE: 0.2930765963078498 km (standarized) * Validation MAE: 0.4236134304387459 km (standarized) * Training MSE: 0.1674885563111346 km (standarized) * Validation MSE: 0.35156953540544217 km (standarized) * Training RMSE: 0.40925365766372157 km (standardized) * Validation RMSE: 0.5929329940266793 km (standardized) | * The model achieves high R^2 values for both the training and validation sets, indicating a strong fit to the data. * The MAE and MSE errors are the lowest among the three models, indicating a smaller average prediction error compared to the other models. * The RMSE error on the validation set is 0.59 km, indicating the average prediction error of the model. |

Overall, the Random Forest model outperforms both the Logistic Regression and Decision Tree models in terms of prediction accuracy, as evidenced by lower MAE, MSE, and RMSE errors. It demonstrates a better ability to capture the underlying patterns and relationships in the data, resulting in improved performance.

**Model Insite from Shapley Plots**





Analyzing the Shapley plots, we can observe the most predictive features for the selected decision tree model in predicting the increase or decrease in diameter.

For increasing diameter prediction, the top three features are:

'n\_obs\_used': This feature shows a positive relationship with the diameter prediction. As the number of observations used increases, it tends to result in a higher diameter prediction.

'albedo': This feature also exhibits a positive impact on the diameter prediction. Higher values of albedo contribute to an increase in the predicted diameter.

'n': The feature 'n' shows a positive relationship with the diameter prediction. An increase in 'n' leads to a higher diameter prediction.

These features indicate that factors such as the number of observations used, albedo, and 'n' have a significant influence on predicting an increase in asteroid diameter.

The features with the lowest values, indicating a strong negative impact on the diameter prediction, are:

'first\_month\_obs': This feature has a negative relationship with the diameter prediction. A lower value of 'first\_month\_obs' is associated with a decrease in the predicted diameter.

'orbit\_id': This feature also shows a negative impact on the diameter prediction. Lower values of 'orbit\_id' contribute to a decrease in the predicted diameter.

'last\_obs\_year': Similarly, 'last\_obs\_year' exhibits a negative relationship with the diameter prediction. A lower value of 'last\_obs\_year' is associated with a decrease in the predicted diameter.

These features suggest that factors such as the first month of observation, the specific orbit ID, and the last year of observation have a strong influence on predicting a decrease in asteroid diameter.

**References**

* Basu, V. (2019, 24 April). Prediction of asteroid diameter with the help of multi-layer perceptron regressor. International Journal of Advances in Electronics and Computer Science, 6(4), 2394– 2835. Retrieved from http://www.iraj.in/journal/journal\_file/journal\_pdf/12-555- 156136953136-40.pdf
* BBVA (2018, 7 December). How an asteroid-detection algorithm paved the way for a career in data science for me [Web log post]. Retrieved from https://www.bbva.com/en/how-an-asteroiddetection-algorithm-paved-the-way-for-a-career-in-data-science-for-me/
* Blakelobato (2020a, 9 January). Predicting asteroid diameter [Electronic resource: Python source code]. GitHub. Retrieved from https://github.com/blakelobato/Predicting-AsteroidDiameter/blob/master/Main\_Analysis\_Notebook.ipynb
* Blakelobato (2020b, 9 January). Predicting asteroid diameter [Electronic resource: Data]. GitHub. Retrieved from https://github.com/blakelobato/Predicting-Asteroid-DiameterDash/blob/master/model/Pred\_Ast\_Diam\_2.csv
* Blakelobato (2020c, 10 January). Shapley plot interactions [Electronic resource: data]. GitHub. Retrieved from https://github.com/blakelobato/Predicting-AsteroidDiameter/blob/master/SHAPELY\_PLOTS.ipynb
* Blakelobato (2020d, 10 January). Predicting asteroid’s diameter using machine learning [Web log post]. Medium. Retrieved from https://medium.com/swlh/predicting-asteroids-diameterusing-machine-learning-e1da883c2196